ORIENTATION RECONSTRUCTION ALGORITHMS FOR X-RAY SERIAL DIFFRACTION DATA

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Ti-Yen Lan

August 2018

ORIENTATION RECONSTRUCTION ALGORITHMS FOR X-RAY SERIAL DIFFRACTION DATA

Ti-Yen Lan, Ph.D.

Cornell University 2018

Due to irreversible radiation damage, structure determination of biological macromolecules using X-rays is often done by taking snapshots from individual copies of the sample and assembling the snapshots in the end to solve the 3D structures. It is difficult to control the orientations of micron or sub-micron sized specimens when delivered to the X-ray beam. Furthermore, the signals in the snapshots may be so weak that each of them cannot be oriented separately.

This thesis develops algorithms to address the task of 3D reconstruction from unoriented, noisy snapshots, with special focus on two X-ray methods. For the first one, single particle imagining at X-ray free electron lasers, we discuss the difficulty of orientation reconstruction of samples through computer simulation, and then present the analysis results of two experimental datasets. For the second technique, serial microcrystallography at synchrotron storage ring sources, we first describe the development of our reconstruction algorithm through two proof-of-concept studies. In these studies, diffraction patterns were collected from large protein crystals to simulate the signal level of those collected from protein microcrystals at storage ring sources. Finally, we demonstrate our method by solving a protein structure from microcrystal diffraction patterns collected at a storage ring synchrotron source. These data would have been discarded by crystallographers because of their weak signals. Through the detailed presentation of the analysis processes, this thesis is also meant to be a self-contained tutorial on reconstruction problems using X-ray sources.

BIOGRAPHICAL SKETCH

Ti-Yen Lan was born in 1990 in Taoyuan, Taiwan, where he lived until 18 years old. He then attended National Taiwan University and received his Bachelor of Science degree in physics in 2012. After a year-long military service in Hualien, Taiwan, a place known for its magnificent gorges, he moved to Ithaca, a beautiful town also full of gorges, for his doctoral studies in physics at Cornell University.

To Patty and Evelyn.

ACKNOWLEDGEMENTS

I would like to first express my gratitude to my advisor, Veit Elser, for his guidance throughout the past five years. Working with Veit has been a lot of fun. I was continuously challenged to seek simpler solutions to problems because of his appreciation of simplicity. This training is invaluable in shaping my view as a scientist. Besides guidance in research, his advice on career and life helped me navigate through many stressful situations. It would be ungrateful to ask for a better advisor than Veit.

A good portion of the work in this thesis was impossible without the constant help and encouragement from Sol Gruner and his group. Sol was like my second advisor. My discussions with him have not only enriched my knowledge about experiments, but also benefited my research in algorithm development. His patience, humbleness and passion about research exemplify the qualities of a good scientist for me.

I also want to thank Tomas Arias, who served on my committee, for bringing the interesting question about Bayes' rule in my A-exam. What I have learned from answering that question remains very helpful to my research.

Discussions with the senior Elser group members, Duane Loh, Kartik Ayyer, Zhen Wah Tan, Hyung Joo Park and Yi Jiang, have stimulated many interesting ideas for my research. I especially thank Duane Loh, Kartik Ayyer and Yi Jiang for providing me useful career suggestions. Outside the Elser group, I want to thank my amazing collaborators in the Gruner group, Jeney Wierman, Mark Tate and Hugh Philipp, for bringing me high-quality data. I thank Jeney, also, for much of the crystallographic structure analyses. Without their efforts, I was not able to do my work. Their patience in answering my naive questions about experiments is also appreciated.

Barry Robinson, the IT manager at LASSP, has been very kind in helping me with computing issues. Barry is always open for discussions whenever I show up with questions to bug him. Besides Veit, he might be the person who has taught me the most

about computation.

I could not survive the past five years in Ithaca without the company of my friends Archishman Raju, Hao Shi, Phil Burnham, Andre Frankenthal, Katherine Quinn and Brendan Faeth. The after-lunch coffee times we spent at Gimme Coffee will remain as an important part of my memory about graduate school.

I would like to thank my parents and my brother for their constant support. Finally, I thank my wife, Patty, for always being there. It has been a beautiful adventure to be with you and our adorable Evelyn.

This work was supported by the Department of Energy (DOE) grants DE-FG02-11ER16210 and DE-SC0005827, and the Taiwan Government Scholarship to Study Abroad. As a theorist, my role in the work presented here was confined to discussion of the acquisition and/or analysis of the data collected by my experimental collaborators. This would not have been possible without the sources of support that my collaborators used to acquire these data: The Gruner group was supported by the DOE grants DE-FG02-10ER46693, DE-SC0016035 and DE-SC0017631. Use of the Cornell High Energy Synchrotron Source (CHESS) was supported by the NSF award DMR-1332208, and the Macromolecular Diffraction at CHESS (MacCHESS) resource was supported by the NIGMS award GM-103485. Use of the Linac Coherent Light Source (LCLS), SLAC National Accelerator Laboratory, was supported by the U.S. DOE, Office of Science, Office of Basic Energy Sciences under Contract No. DE-AC02-76SF00515. Use of the Advanced Photon Source, a U.S. DOE Office of Science User Facility operated for the DOE Office of Science by Argonne National Laboratory, was supported under Contract No. DE-AC02-06CH11357.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Structure determination of biological macromolecules is practically a battle against structural damage caused by photons or electrons. For the past few decades, crystallography has been the method of choice because the periodic arrangement of structural units enhances the weak signals of individual molecules through constructive interference of the scattered waves, which produces the sharp Bragg peaks in the recorded diffraction patterns. The signal enhancement allows the collection of adequate information to resolve the structure of the constituent molecules before their structures are compromised by radiation or electron damage. With the developments in X-ray synchrotron sources, experimental technology and data analysis methods, crystallography has contributed over 126,000 structures to the Protein Data Bank (PDB) to date. What challenges crystallography, however, is to form sufficiently large single crystals that diffract to high resolution and minimize the irreversible structural damage. The structure determination of many functionally important proteins, such as membrane proteins, may fail at this stage.

Another route to the structure solution of macromolecules is through the single-particle approach, which avoids the necessity of crystallization. In the single-particle approach, structural information is collected from many individual macromolecules, or particles, of reasonably similar structures at random orientations. To minimize structural damage, either the net dose is limited or the exposure time of the illumination is made very short. The 3D structure is solved by assembling many noisy signals from the randomly-oriented particles. A representative technique is single-particle cryoelectron microscopy (cryo-EM) [23], which solves the 3D structure by merging 2D projection images collected from individual, randomly-oriented particles. The particles are cryogenically preserved in a thin layer of vitreous ice to mitigate the electron damage. In the past few years,

the advance in direct electron detectors has brought the resolution of this technique to near-atomic level [43], and makes it competitive with crystallography.

Due to the smallness of biological macromolecules, single-particle data is usually extremely noisy, which makes the 3D structure reconstruction from the unoriented, noisy data a daunting task. This thesis focuses on developing analysis methods to tackle the 3D reconstruction problem from unoriented X-ray data. The main applications of our methods lie in single particle imaging (SPI) at X-ray free electron lasers (XFELs) and serial microcrystallography (SMX) at storage ring synchrotron sources, which share the same characteristics that the data frames are too noisy to be oriented on a per frame basis. The structure of the thesis is outlined as follows.

Chapter 2 lays out an overview of the theoretical background of structure determination using X-ray diffraction. The interactions of X-rays with matter are described in the language of scattering theory, which helps to visualize the competition between different types of interactions and to relate the commonly used terms in X-ray crystallography. Subsequently, we present the formalism of how structural information is encoded in diffraction patterns, and elaborate on a special kind of sample: crystals. The missing phase problem in X-ray diffraction measurements and different methods for phase retrieval are also discussed, for both non-crystalline and crystalline cases. After that, we introduce the expand-maximize-compress (EMC) algorithm, the core algorithm in this thesis to reconstruct 3D intensity maps from unoriented diffraction patterns. Finally, we explain variants of the EMC algorithm for different experimental conditions or that save on computational resources.

Chapter 3 describes our contribution to SPI experiments from the theoretical side. We first present a computer simulation study on the selection of appropriate samples for the first few R&D experiments based on the difficulty to assemble the unoriented

diffraction patterns. The analysis results of two SPI datasets are then discussed. Using the first dataset, we introduce a metric derived from Poisson statistics that measures the consistency of a diffraction pattern with a known structure model. We then show the 3D structure of a virus particle solved at a modest resolution from the second dataset.

In Chapter 4, we develop the EMC algorithm through two proof-of-concept studies. In these studies, diffraction patterns were collected from large protein crystals illuminated by a dim lab X-ray source to simulate those collected from many microcrystals. The orientations of the data frames were kept unknown to the reconstruction algorithm. By increasing the experimental complexity, we show that our reconstruction method should be able to undertake the analysis of a real SMX dataset.

Chapter 5 presents a step-by-step analysis of a real SMX dataset collected at a storage ring source. In particular, we demonstrate that 3D intensity reconstruction is still feasible from data frames whose signals are too weak to be considered by crystallographers. Furthermore, the structure solved from our reconstructed Bragg intensities compares favorably with that solved from data with stronger signals using more conventional means. The implementation details of our analysis package is given in Appendix A.

CHAPTER 2

THEORY

This chapter gives an overview of the key theoretical concepts behind this thesis. We first discuss the major interactions of X-rays with matter, and how structural information of materials is encoded in the spatial distribution of coherently scattered X-rays, which is recorded in the form of diffraction patterns by pixelated detectors. Since the diffraction measurements only provide the magnitudes of the scattered waves, the retrieval of the missing phases is described that uses prior information on the sample to solve its structure. When the sample is radiation sensitive, information about its 3D structure may be obtained by collecting diffraction patterns from individual copies of the sample at various orientations. However, it is difficult to experimentally control the orientations of micron or sub-micron sized specimens, and orientation reconstruction is challenging because the diffraction patterns from the small specimens are shot-noise limited. In the last part of the chapter, we introduce the EMC algorithm [46], which assembles the noisy, unoriented diffraction patterns to form the 3D intensity distribution of the sample by maximizing the data likelihood. Variants of the algorithm are discussed that tackle different experimental conditions and ease the computational demands.

2.1 Interactions of X-rays with matter

In the energy range of X-rays (100 eV - 100 keV), the photon interaction cross section of an isolated atom is mainly contributed by the photoelectric effect, coherent scattering and incoherent scattering [36]:

$$\sigma_{\text{tot}} = \sigma_{\text{pe}} + \sigma_{\text{coh}} + \sigma_{\text{incoh}}. \tag{2.1}$$

The details of these interactions and their contributions to X-ray diffraction measurements are as follows.

2.1.1 Photoelectric effect

In the photoelectric effect, an atom absorbs all the energy of the incident photon and ejects a core electron. The resulting vacancy is then filled by an electron from a higher energy level. The energy difference is released by either X-ray fluorescence or ejecting another electron, which is called an Auger electron. The emitted fluorescence photon has a random direction and phase, and contributes to the background in diffraction measurements incoherently.

2.1.2 Coherent scattering

Coherent scattering is the signal of interest in most X-ray diffraction experiments. As suggested by its name, the scattered wave is coherent, and the phase depends on the positions of the scatterers. This section describes the coherent scattering of X-rays by electrons and atoms.

Coherent scattering by electrons

Consider a plane wave of linear polarization $\mathbf{E}_{\text{inc}}$ incident on a particle of charge q_p and mass m placed at the origin. The charged particle undergoes an oscillating acceleration $\mathbf{a} = q_p\mathbf{E}_{\text{inc}}/m$ and emits electromagnetic radiation that is coherent with the incident wave. The electric field of the scattered wave observed at position $\mathbf{r}$ in the far field can

be expressed as

$$\mathbf{E}_{\text{sc}} = \frac{q_p}{4\pi\epsilon_0 c^2}\frac{\mathbf{r}\times(\mathbf{r}\times\mathbf{a})}{|\mathbf{r}|^3} = \frac{q_p^2}{4\pi\epsilon_0 mc^2}\frac{\mathbf{r}\times(\mathbf{r}\times\mathbf{E}_{\text{inc}})}{|\mathbf{r}|^3}, \tag{2.2}$$

where ϵ_0 is the vacuum permittivity, and c is the speed of light.

The scattered intensity, which is defined as the average power transferred per unit solid angle, is given by

$$I_{\text{sc}} = \frac{\epsilon_0 c}{2}|\mathbf{r}|^2|\mathbf{E}_{\text{sc}}|^2 = \sin^2\alpha\left(\frac{q_p^2}{4\pi\epsilon_0 mc^2}\right)^2\langle S\rangle_{\text{inc}} = P\left(\frac{q_p^2}{4\pi\epsilon_0 mc^2}\right)^2\langle S\rangle_{\text{inc}}, \tag{2.3}$$

where α is the angle between $\mathbf{r}$ and $\mathbf{E}_{\text{inc}}$, $\langle S\rangle_{\text{inc}} = \frac{\epsilon_0 c}{2}|\mathbf{E}_{\text{inc}}|^2$ is the average incident energy flux density, and $P = \sin^2\alpha$ is called the polarization factor. From Equation (2.3) we define the differential cross section for the coherent scattering from a charged particle as

$$\left(\frac{d\sigma}{d\Omega}\right)_{\text{coh}} = \frac{I_{\text{sc}}}{\langle S\rangle_{\text{inc}}} = P\left(\frac{q_p^2}{4\pi\epsilon_0 mc^2}\right)^2. \tag{2.4}$$

The inverse proportionality to the squared mass suggests that electrons are the dominant scatterers in coherent scattering. The quantity

$$r_e = \frac{e^2}{4\pi\epsilon_0 m_e c^2} \sim 2.82\times10^{-15}\ \text{m}, \tag{2.5}$$

which has units of length, is called the classical electron radius. Here e and m_e represent the electron charge and mass, respectively.

Coherent scattering by atoms

The coherent X-rays are mainly scattered by the electron cloud of an atom because the heavy nucleus is barely moved by the electric field of the incident wave. For an atom with electron density $\rho(\mathbf{x})$ placed at the origin, the scattered wave observed at position $\mathbf{r}$ in the far field can be approximated by the superposition of the waves scattered by the

individual electrons:

$$\mathbf{E}_{\text{sc}} = r_e \, \frac{\mathbf{r} \times (\mathbf{r} \times \mathbf{E}_{\text{inc}})}{|\mathbf{r}|^3} \int d\mathbf{x} \; \rho(\mathbf{x}) \; e^{-i\mathbf{q}\cdot\mathbf{x}}, \tag{2.6}$$

where $\mathbf{q}$ is the wave-vector difference between the scattered and incident waves. Comparing Equation (2.6) with Equation (2.2), we can see that the scattering amplitude of an isolated atom is quantified by the integral

$$f(\mathbf{q}) = \int d\mathbf{x} \; \rho(\mathbf{x}) \; e^{-i\mathbf{q}\cdot\mathbf{x}}, \tag{2.7}$$

which is called the atomic scattering factor. The coherent scattering cross section of the atom can therefore be written as

$$\left(\frac{d\sigma}{d\Omega}\right)_{\text{coh}} = P \; r_e^2 \; |f(\mathbf{q})|^2. \tag{2.8}$$

The above derivation of the atomic scattering factor ignores the internal structure of an atom — the electron energy levels. When the incident photons have energy close to an absorption edge to excite a core electron, the atomic scattering factor is corrected by

$$f(\mathbf{q}, \lambda) = f(\mathbf{q}) + f'(\lambda) + i f''(\lambda), \tag{2.9}$$

where $f'(\lambda)$ and $f''(\lambda)$ are called the anomalous scattering factors and are exploited to gain phase information in crystallography. The X-ray energy considered in this thesis is assumed to be far from any absorption edge, so Equation (2.7) is a good approximation of the atomic scattering factor.

2.1.3 Incoherent scattering

In contrast to coherent scattering, the incident photons lose a fraction of the energy to the electrons of an atom in incoherent scattering. This process, also known as Compton

scattering, is best described by the elastic collision of a photon with an electron. A photon of wavelength λ has momentum h/λ, where h is the Planck constant. When the photon strikes on an atomic electron, approximated as being at rest, the electron recoils and emits another photon of wavelength λ' at scattering angle θ. We can determine the wavelength difference by energy and momentum conservation as

$$\lambda' - \lambda = \frac{h}{m_e c} (1 - \cos\theta). \tag{2.10}$$

The quantity $h/m_e c$ is called the Compton wavelength and has the numerical value 2.43×10^{-2} Å. The small wavelength difference makes Compton photons unresolvable from coherent photons by normal X-ray detectors.

By neglecting the exchange interactions between electrons of an atom, the differential scattering cross section of Compton scattering can be approximated as

$$\left(\frac{d\sigma}{d\Omega}\right)_{\text{incoh}} = P\, r_e^2 \left(Z - \frac{|f(\mathbf{q})|^2}{Z}\right), \tag{2.11}$$

where Z is the atomic number of the atom [31]. Because $|f(\mathbf{q})|$ drops rapidly from Z to 0 with the increase of $|\mathbf{q}|$, Compton scattering can be ignored at small scattering angles and becomes significant only at large scattering angles. As we will see in the next section, the periodic arrangement of the atoms in a crystal can enhance the coherent scattering signal by coherently adding up the scattering amplitudes of the atoms. On the other hand, it is the atomic scattering intensities, not amplitudes, that add up in Compton scattering because the scattered photons are incoherent. As a result, incoherent scattering is insignificant for X-rays scattering from crystals.

2.2 X-ray diffraction basics

Here we describe how structures can be obtained from X-ray diffraction measurements. The general and a special cases of samples — crystals — are discussed.

2.2.1 First-order Born approximation

Consider a sample of electron density $\rho(\mathbf{x}) = \sum_j \rho_j(\mathbf{x} - \mathbf{x}_j)$, with j denoting the individual atoms. When a plane electromagnetic wave illuminates the sample, the scattered wave in the far field is the superposition of the emitted radiation driven by the total electric field at the position of each atom. The first-order Born approximation replaces the total electric field by the electric field of the incident wave, so the electric field of the scattered wave at the far-field position $\mathbf{r}$ can be written as

$$\begin{aligned}\mathbf{E}_{\text{sc}} &= r_e \, \frac{\mathbf{r} \times (\mathbf{r} \times \mathbf{E}_{\text{inc}})}{|\mathbf{r}|^3} \int d\mathbf{x} \, \rho(\mathbf{x}) \, e^{-i\mathbf{q}\cdot\mathbf{x}} \\ &= r_e \, \frac{\mathbf{r} \times (\mathbf{r} \times \mathbf{E}_{\text{inc}})}{|\mathbf{r}|^3} \sum_j e^{-i\mathbf{q}\cdot\mathbf{x}_j} \int d\mathbf{x} \, \rho_j(\mathbf{x}) \, e^{-i\mathbf{q}\cdot\mathbf{x}} \\ &= r_e \, \frac{\mathbf{r} \times (\mathbf{r} \times \mathbf{E}_{\text{inc}})}{|\mathbf{r}|^3} \sum_j f_j(\mathbf{q}) \, e^{-i\mathbf{q}\cdot\mathbf{x}_j}. \end{aligned} \tag{2.12}$$

Multiple scattering in the sample is assumed to be negligible in the first-order Born approximation. This assumption applies to optically thin samples, where the phase change due to the sample can be ignored [72].

2.2.2 X-ray diffraction of materials

From Equation (2.12), we readily obtain the scattered intensity of the sample:

$$I_{sc} = P \, r_e^2 \, \langle S \rangle_{\text{inc}} \left| \sum_j f_j(\mathbf{q}) \, e^{-i\mathbf{q}\cdot\mathbf{x}_j} \right|^2. \tag{2.13}$$

When the incident X-ray energy is far from any absorption edge of the constituent atoms, the atomic scattering factors, $f_j(\mathbf{q})$, are given by Equation (2.7). With the Fourier transform of $\rho(\mathbf{x})$ given by

$$\hat{\rho}(\mathbf{q}) = \int d\mathbf{x}\ \rho(\mathbf{x})\ e^{-i\mathbf{q}\cdot\mathbf{x}} = \sum_j f_j(\mathbf{q})\ e^{-i\mathbf{q}\cdot\mathbf{x}_j}, \tag{2.14}$$

the Fourier magnitudes have inversion symmetry:

$$|\hat{\rho}(\mathbf{q})| = |\hat{\rho}(-\mathbf{q})|, \tag{2.15}$$

also known as the Friedel symmetry. The scattered X-rays are recorded in the form of diffraction patterns by a pixelated detector. The mean photon number, $\langle K_i \rangle$, recorded by pixel i over exposure time Δt is given by

$$\langle K_i \rangle = P_i\ r_e^2 J_{\text{inc}} |\hat{\rho}(\mathbf{q}_i)|^2 \Delta t\ \Delta\Omega_i, \tag{2.16}$$

where P_i is the polarization factor for pixel i, J_{inc} is the average incident photon flux density, $\mathbf{q}_i$ is the wave-vector difference between the wave scattered to pixel i and the incident wave, and $\Delta\Omega_i$ is the solid angle subtended by pixel i.

The wave-vector difference, $\mathbf{q}$, is also called the spatial frequency, and the space of spatial frequencies is usually referred to as reciprocal space by crystallographers. The spatial frequency magnitude is given by $|\mathbf{q}| = 4\pi \sin(\theta/2)/\lambda$, where λ denotes the wavelength of the incident wave, and θ is the angle between the incident wave vector and the scattered wave vector. By defining the time-integrated intensity as

$$W(\mathbf{q}) = r_e^2 J_{\text{inc}} |\hat{\rho}(\mathbf{q})|^2 \Delta t, \tag{2.17}$$

the mean photon number, $\langle K_i \rangle$, recorded by pixel i is given by

$$\langle K_i \rangle = P_i\ W(\mathbf{q}_i)\ \Delta\Omega_i. \tag{2.18}$$

Since the scattered waves have the same wavelength as the incident wave, the spatial frequencies associated with the pixels all lie on a sphere in reciprocal space. This sphere,

called the Ewald sphere, has radius $2\pi/\lambda$ and is centered at $\mathbf{q} = -\mathbf{k}_{\text{inc}}$, where $\mathbf{k}_{\text{inc}}$ is the incident wave vector. As a result, each measured diffraction pattern corresponds to an Ewald-sphere slice of the 3D contrast, $W(\mathbf{q})$, multiplied by the pixel-wise polarization factors and solid angles.

2.2.3 X-ray diffraction of crystals

A crystal features the periodic arrangement of a repeating structural unit, also known as the unit cell. In particular, the electron density of a crystal can be expressed by

$$\rho_c(\mathbf{x}) = \sum_{\mathbf{y}\in S} \rho(\mathbf{x} - \mathbf{y}), \tag{2.19}$$

where S is a finite set of translation vectors and $\rho(\mathbf{x})$ is the electron density of the molecules in a unit cell. The crystal parameters are defined by a lattice $\Lambda \subset \mathbb{R}^D$, and $S \subset \Lambda$ is in practice a very large and compact subset.

X-ray diffraction measurements provide information on the Fourier magnitudes, $|\hat{\rho}_c(\mathbf{q})|$, of the crystal, where

$$\begin{aligned}\hat{\rho}_c(\mathbf{q}) &= \int d\mathbf{x}\ \rho_c(\mathbf{x})\ e^{-i\mathbf{q}\cdot\mathbf{x}} \\ &= \sum_{\mathbf{y}\in S} e^{-i\mathbf{q}\cdot\mathbf{y}} \int d\mathbf{x}\ \rho(\mathbf{x})\ e^{-i\mathbf{q}\cdot\mathbf{x}} \\ &= \hat{s}(\mathbf{q})\ \hat{\rho}(\mathbf{q}).\end{aligned} \tag{2.20}$$

Here $\hat{\rho}(\mathbf{q})$ is the Fourier transform of $\rho(\mathbf{x})$, and $\hat{s}(\mathbf{q})$ is a modulating function that depends on the crystal size. When the size of S grows, the values of $\hat{s}(\mathbf{q})$ increasingly concentrate on the reciprocal lattice points, $\mathbf{Q} \in \Lambda^*$, where Λ^* is the dual lattice to Λ in reciprocal space. This concentration of diffracting power leads to the so-called Bragg peaks in reciprocal space.

Crystals often have symmetries other than the translational symmetries defined by the lattice, Λ. For an (idealized) infinite crystal, the electron density, $\rho_c(\mathbf{x})$, is unchanged by elements in a finite group G. An element $g \in G$ acts on the density function, $\rho_c(\mathbf{x})$, by the composition of an orthogonal matrix transformation (rotation or reflection), R_g, and a translation, T_g:

$$g(\rho_c(\mathbf{x})) = \rho_c(R_g \cdot \mathbf{x} + T_g). \tag{2.21}$$

Thus in addition to

$$\rho_c(\mathbf{x}) = \rho_c(\mathbf{x} + \mathbf{y}), \quad \mathbf{y} \in \Lambda, \tag{2.22}$$

the density function also satisfies

$$\rho_c(\mathbf{x}) = g(\rho_c(\mathbf{x})), \quad g \in G. \tag{2.23}$$

The set of orthogonal matrices R_g identifies G with a point group (transformations that fix the origin), while the set of pairs (R_g, T_g) together with the group of lattice translations, Λ, specify the crystal's space group. The space group manifests itself by the rotational symmetry and systematic extinctions of Bragg peaks in reciprocal space [30].

From Equation (2.16), the mean photon number measured by pixel i is given by

$$\begin{aligned} \langle K_i \rangle &= P_i \, r_e^2 J_{\text{inc}} \, |\hat{\rho}_c(\mathbf{q}_i)|^2 \Delta t \, \Delta\Omega_i \\ &= P_i \, r_e^2 J_{\text{inc}} \, |\hat{s}(\mathbf{q}_i)|^2 \, |\hat{\rho}(\mathbf{q}_i)|^2 \Delta t \, \Delta\Omega_i. \end{aligned} \tag{2.24}$$

An interesting observation about $|\hat{s}(\mathbf{q})|^2$ is that it is periodic over the reciprocal lattice. Assume that N_c is the number of unit cells in the crystal. At the Bragg positions, $|\hat{s}(\mathbf{Q})|^2 = N_c^2$, and the integration of $|\hat{s}(\mathbf{q})|^2$ over a Bragg peak equals N_c. For sufficiently large crystals (at least several tens of unit cells in each dimension), we can approximate $|\hat{s}(\mathbf{q})|^2$ by a sum of Dirac delta functions:

$$|\hat{s}(\mathbf{q})|^2 \approx N_c \sum_{\mathbf{Q} \in \Lambda^*} \delta(\mathbf{q} - \mathbf{Q}). \tag{2.25}$$

The integral of the measured photons over a Bragg peak at $\mathbf{Q}$ hence gives information on $|\hat{\rho}(\mathbf{Q})|^2$, and the signal strength is proportional to the number of unit cells in the crystal, or equivalently, the crystal volume.

For the crystals considered in this thesis, protein crystals, the alignment of the unit cells is imperfect. Instead, a protein crystal consists of many slightly misaligned domains, called the mosaic blocks. Each mosaic block diffracts X-rays at a slightly different orientation, which results in the broadening of Bragg peaks in reciprocal space. Nevertheless, the widths of the Bragg peaks are still small enough so that the function $|\hat{\rho}(\mathbf{q})|^2$ can be approximated as a constant over each peak. The integrated value over a Bragg peak at $\mathbf{Q}$ is again proportional to $|\hat{\rho}(\mathbf{Q})|^2$ and the crystal volume.

2.2.4 Phase problem

In order to reconstruct the electron density of the sample, $\rho(\mathbf{x})$, we need the magnitudes and phases of its Fourier transform, $\hat{\rho}(\mathbf{q})$. However, X-ray diffraction measurements only provide the Fourier magnitudes, $|\hat{\rho}(\mathbf{q})|$. In addition, experimental limitations make some values of $|\hat{\rho}(\mathbf{q})|$ inaccessible. For example, the value of $|\hat{\rho}(\mathbf{0})|$ cannot be measured because scattered photons with zero spatial frequency are indistinguishable from the unscattered photons. Both forms of information loss should be compensated by other sources of information, and this is the task of phase retrieval.

Phasing crystallographic data

Consider a 1D discrete periodic function, $f(x_n)$, which has period ℓ and sample points $x_n = n\Delta x = n\ell/N$, $n = 0, 1, 2, \ldots, N-1$. This function can be fully represented by a

Fourier series:

$$f(x_n) = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} \hat{f}(q_k)\, e^{iq_k x_n}, \tag{2.26}$$

where $q_k = k\Delta q = k(2\pi/\ell)$. Crystallographic measurements give the absolute values of the Fourier components, $|\hat{f}(q_k)|$, so the reconstruction of the discrete signal, $f(x_n)$, is under-constrained by a factor of 2.

Methods for phasing crystallographic data can be roughly divided into four categories: isomorphous replacement, anomalous dispersion, molecular replacement and direct methods [63]. In isomorphous replacement, phases are calculated from the differences in Fourier magnitudes between a native crystal and its heavy-atom derivatives, assuming that the addition of heavy atoms does not change the original crystal structure. The method of anomalous dispersion takes advantage of the anomalous scattering factors (Equation (2.9)) of the heavy atoms present in a crystal by tuning the incident X-ray energy close to the corresponding absorption edges. Since the complex anomalous scattering factors, $f'(\lambda) + if''(\lambda)$, are independent of the spatial frequency, $\mathbf{q}$, the Friedel symmetry (Equation (2.15)) is broken, and the differences between the Fourier magnitudes of the Friedel pairs offer extra information for phase determination.

When a reasonably large fraction of the contrast in a crystal is known, molecular replacement can be used to estimate the phases. The known structure is oriented and translated to fit its autocorrelation function with that derived from the experimental data, from which one can derive the phases that are hopefully close to the true values. Direct methods use prior knowledge on the unknown structure to constrain the phase solution, for example, the sparsity and non-negativity of the signals or the phase relations between certain sets of Fourier components. The success of direct methods usually requires atomic resolution data. The relation between data quality and the hardness of phase retrieval was recently studied in Ref. [18].

Phasing for aperiodic samples

The theoretical foundation of phase retrieval for non-crystalline samples traces back to an observation by David Sayre in 1952: structure determination for isolated objects would be possible if the intensity measurement could be sufficiently oversampled [64]. This idea has spawned the technique of coherent X-ray diffraction imaging (CXDI), where phases are retrieved from the oversampled intensity measurement and the prior knowledge of the sample size and shape.

Consider a 1D band-limited signal, $f(x)$, which is non-zero in the interval $x \in (0, \ell)$ and zero elsewhere. From Shannon's sampling theorem [65], $f(x)$ can be fully represented without aliasing by its Fourier components, $\hat{f}(q_k)$:

$$f(x) = \frac{\sqrt{2\pi}}{\ell} \sum_{k=-\infty}^{\infty} \hat{f}(q_k)\, e^{iq_k x}, \tag{2.27}$$

where $q_k = k\Delta q = k(2\pi/\ell)$. For an X-ray diffraction measurement with object size L, the Fourier components, $\hat{f}(q_k)$, correspond to the Fourier intensities, $|\hat{\rho}(\mathbf{q})|^2$, and the signal, $f(x)$, corresponds to the inverse Fourier transform of $|\hat{\rho}(\mathbf{q})|^2$, the signal autocorrelation

$$a(\mathbf{x}) = a(-\mathbf{x}) = \int d\mathbf{x}'\, \rho(\mathbf{x}')\rho(\mathbf{x}' + \mathbf{x}), \tag{2.28}$$

which has band limit, or support size, $\ell = 2L$. Therefore, $a(\mathbf{x})$ can be uniquely represented if the intensity measurement is sampled at a rate finer than

$$\Delta q = \frac{\pi}{L}, \tag{2.29}$$

and the electron density, $\rho(\mathbf{x})$, is reconstructed given $a(\mathbf{x})$ as well as prior information on $\rho(\mathbf{x})$.

The difficulty of a phase retrieval problem can be further quantified by the ratio of the number of constraints provided by the signal autocorrelation to the number of free

variables in the signal [19]. This quantity — the constraint ratio — is defined as

$$\Omega = \frac{1}{2}\frac{A_{\text{auto}}}{A_S}, \tag{2.30}$$

where A_{auto} and A_S denote the support sizes of $a(\mathbf{x})$ and $\rho(\mathbf{x})$, respectively, and the factor of $1/2$ is due to the centrosymmetry of $a(\mathbf{x})$. When $\Omega > 1$, the reconstruction is possible without any additional information. In dimensions higher than 1D, Ω is always no less than 2 as long as the intensity measurements are oversampled at a rate finer than Δq defined in Equation (2.29). The definition of Ω immediately suggests: (1) The oversampling condition in Equation (2.29) can be slightly relaxed as long as $A_{\text{auto}} > 2A_S$ to constrain the phase retrieval problem. (2) Increasing the oversampling rate higher than Δq does not provide extra information. Finally, we note that $\Omega = 1/2$ for crystals since the signal autocorrelation has the same support size as the unit cell. This again shows that phase retrieval for crystallographic data is under-constrained by a factor of 2.

The modern phase retrieval algorithms commonly used in CXDI are close descendants of the hybrid-input-output (HIO) algorithm [22] and can be generalized by the difference map algorithm [17]. This class of algorithms searches for solutions that satisfy two sets of constraints, which are enforced by the projections, $P_1(\rho)$ and $P_2(\rho)$. In the context of CXDI, $P_1(\rho)$ modifies the vector variable, ρ, through the Fourier synthesis using the measured Fourier intensities, $I(\mathbf{q})$:

$$P_1(\rho) = \mathcal{F}^{-1}\{\tilde{\rho}\}, \tag{2.31}$$

where

$$\tilde{\rho}(\mathbf{q}) = \begin{cases} \sqrt{I(\mathbf{q})}\,\dfrac{\mathcal{F}\{\rho\}(\mathbf{q})}{|\mathcal{F}\{\rho\}(\mathbf{q})|}, & \text{if } I(\mathbf{q}) \text{ is measured} \\ \mathcal{F}\{\rho\}(\mathbf{q}), & \text{otherwise} \end{cases}. \tag{2.32}$$

Here $\mathcal{F}$ and $\mathcal{F}^{-1}$ denote the fast Fourier transform (FFT) and its inverse. The other

projection, $P_2(\rho)$, enforces the support and non-negativity constraints:

$$P_2(\rho) = \begin{cases} \rho(\mathbf{x}), & \text{if } \mathbf{x} \in S \text{ and } \rho(\mathbf{x}) > 0 \\ 0, & \text{otherwise} \end{cases}, \tag{2.33}$$

where S represents the known support of the isolated sample.

The solution search in the difference map algorithm is done iteratively through the mapping:

$$\rho \rightarrow \rho' = \rho + \beta \Big(P_1(f_2(\rho)) - P_2(f_1(\rho)) \Big), \tag{2.34}$$

where

$$f_2(\rho) = P_2(\rho) + \beta^{-1}(P_2(\rho) - \rho) \tag{2.35}$$

$$f_1(\rho) = P_1(\rho) - \beta^{-1}(P_1(\rho) - \rho). \tag{2.36}$$

The parameter, β, is usually set as 1 (or -1), but its optimal value should be determined through experimentation. The convergence of the search is monitored by the norm of the difference between the two projections:

$$\Delta = ||P_1(f_2(\rho)) - P_2(f_1(\rho))||_2. \tag{2.37}$$

When the search converges to a fixed point, ρ^*, we have $\Delta \approx 0$, and the electron density of the sample is given by either $P_1(f_2(\rho^*))$ or $P_2(f_1(\rho^*))$, a common element of the two constraint sets.

2.3 EMC algorithm

As discussed in Section 2.2, a measured diffraction pattern gives information on an Ewald-sphere slice of the 3D Fourier intensities, $|\hat{\rho}(\mathbf{q})|^2$. In order to resolve the 3D structure of the sample, multiple diffraction patterns have to be recorded at different sample orientations and merged in reciprocal space. In the case of tomography, the

sample is rotated about a defined axis, and the recorded diffraction patterns can be assembled in reciprocal space at the known rotation angles. However, radiation damage sets a limit on the maximum tolerable dose, $D_{\max}$, of the sample. For frozen-hydrated biological samples, Howells and coworkers show that $D_{\max}$ is proportional to the resolution (a length), while the needed dose scales with the inverse fourth power of the resolution [35]. This limit precludes the collection of multiple diffraction patterns from a single small biological particle such as viruses or protein microcrystals.

If many identical copies of a biological particle are available[1], radiation dose can be distributed by taking just one snapshot of each copy at some particle orientation under the safe dose. The diffraction patterns are subsequently assembled to form the 3D Fourier intensities, $|\hat{\rho}(\mathbf{q})|^2$. This idea of 'single-particle analysis' is the basis for single-particle cryoEM [23], single particle X-ray imaging [54] and serial crystallography [11].

The smallness of the biological particles makes it challenging to control the particle orientations relative to the X-ray beam. Since the number of diffracted photons is proportional to the total number of electrons in a particle, the resulting diffraction patterns are expected to be noisy. The EMC algorithm [46] is designed to assemble the noisy, unoriented diffraction patterns in reciprocal space and reconstruct the 3D Fourier intensities, $|\hat{\rho}(\mathbf{q})|^2$. The following sections describe the details of the algorithm and discuss several variants for different experimental conditions.

2.3.1 Standard EMC algorithm

Given a set of noisy diffraction patterns, K, with unmeasured particle orientations, Ω, the EMC algorithm seeks to construct a consistent 3D intensity model, W, by itera-

[1]For protein microcrystals, this assumption means the constituent protein molecules have the same conformations and the unit cell parameters are identical across all the crystals.

tively maximizing the data likelihood function, $p(K|W)$. However, the maximization of $p(K|W)$ is usually intractable due to the missing information of Ω. By contrast, maximizing the complete likelihood function, $p(K, \Omega|W)$, is more straightforward, but requires knowing the particle orientation in each data frame. This observation motivates a way to reconstruct W by alternately updating the estimates of W and Ω by fixing the values of one or the other until convergence.

The expectation-maximization algorithm [15] offers an explicit formalism to update W by iteratively maximizing an expected log-likelihood function

$$Q(W') = \sum_K \int d\Omega \; p(\Omega|K, W) \log p(K, \Omega|W'). \tag{2.38}$$

It can be shown that

$$\log p(K|W') - \log p(K|W) \geq Q(W') - Q(W), \tag{2.39}$$

so the data likelihood function is guaranteed to be non-decreasing by maximizing $Q(W')$. In the context of the EMC algorithm, expectation maximization represents the update rule on the intensity model: $W \rightarrow W'$.

Consider an intensity reconstruction problem with M_{data} data frames collected from individual biological particles at random orientation. Each data frame, k, measures M_{pix} discrete photon counts, K_{ik}, $i = 1, 2, \ldots, M_{\text{pix}}$. The photon counts are assumed to be sampled from Poisson distributions. The rotations are sampled by the 600-cell subdivision method [46], with the sampling rate specified by the order, $n = 1, 2, \ldots$. The angular resolution is given by $\delta\theta = 0.944/n$, and $M_{\text{rot}} = 10(5n^3 + n)$ denotes the number of discrete rotation samples (labeled by j). Let $W(\mathbf{q})$ be the time-integrated 3D intensity defined by Equation (2.17), where $\mathbf{q}$ represents the spatial frequencies. The EMC algorithm iteratively reconstructs $W(\mathbf{q})$ to be consistent with the data frames.

Each iteration of the EMC algorithm consists of three steps: expand (E), maximize

(M) and compress (C). The E-step calculates the mean photon numbers measured by the pixels given the current 3D intensity estimate, W, at different orientations. When the biological particle has orientation j, the intensity value sampled by pixel i is given by linear interpolation

$$W_{ij} = \sum_{\mathbf{p}} f(\mathbf{p} - \mathbf{R}_j \cdot \mathbf{q}_i) W(\mathbf{p}), \tag{2.40}$$

where $f(\cdot)$ is the interpolation weight, $\mathbf{p}$ denotes the 3D grid points in reciprocal space, $\mathbf{R}_j$ is the rotation matrix that brings the lab frame to the particle reference frame at particle orientation j, and $\mathbf{q}_i$ is the spatial frequency associated with pixel i in the lab frame. In the original EMC paper, the pixel-wise polarization factors and solid angles are assumed to be constant at fixed $|\mathbf{q}|$, and they are absorbed into the definition of $W(\mathbf{q})$. Therefore, W_{ij} represents the mean photon number measured by pixel i at particle orientation j.

The M-step updates the tomographic representation, W_{ij}, of the 3D intensity model by maximizing the expected log-likelihood function, $Q(W')$, defined by Equation (2.38). The definition of $Q(W')$ assigns a provisional probability of orientations conditional on the current intensity model, $p(\Omega|K, W)$, to the complete log-likelihood function, $\log p(K, \Omega|W')$, for each data frame. From Bayes' rule, $p(\Omega|K, W)$ can be expressed by

$$p(\Omega|K, W) = \frac{p(K|\Omega, W)p(\Omega|W)}{\int d\Omega\, p(K|\Omega, W)p(\Omega|W)}, \tag{2.41}$$

which is the normalized likelihood function $p(K|\Omega, W)$, weighted by a prior orientation distribution $p(\Omega|W)$. In the implementation of the EMC algorithm, $p(\Omega|K, W)$ and $p(K|\Omega, W)$ have the discrete representations $P_{jk}(W)$ and $R_{jk}(W)$ for data frame k, respectively. The probability $R_{jk}(W)$ is the product of the Poisson probabilities of the photon count measured by each detector pixel:

$$R_{jk}(W) = \prod_i \frac{W_{ij}^{K_{ik}} \exp(-W_{ij})}{K_{ik}!}. \tag{2.42}$$

Since the particle orientations are generally assumed to be uniformly distributed, we represent the prior orientation distribution, $p(\Omega|W)$, by w_j, which is the fraction of the continuous rotation group assigned to rotation sample j. Finally, the conditional probability $P_{jk}(W)$ is given by

$$P_{jk}(W) = \frac{w_j R_{jk}(W)}{\sum_{j'} w_{j'} R_{j'k}(W)}. \tag{2.43}$$

The expected log-likelihood function defined in Equation (2.38) can be rewritten as

$$Q(W') = \sum_K \int d\Omega \left(p(\Omega|K, W) \log p(K|\Omega, W') + p(\Omega|K, W) \log p(\Omega|W') \right). \tag{2.44}$$

Since the prior orientation distribution, $p(\Omega|W')$, is generally independent of the intensity model, W', the second term in Equation (2.44) can be neglected. In the representation of discrete variables, we have

$$Q(W') = \sum_j \sum_k P_{jk}(W) \left(\sum_i K_{ik} \log W'_{ij} - W'_{ij} \right), \tag{2.45}$$

where an irrelevant constant is again neglected. Maximizing $Q(W')$ with respect to W'_{ij}, we obtain the update rule

$$W_{ij} \to W'_{ij} = \frac{\sum_k P_{jk}(W) K_{ik}}{\sum_k P_{jk}(W)}, \tag{2.46}$$

which can be interpreted as the average of the photon counts, K_{ik}, weighted by the conditional probabilities, $P_{jk}(W)$, over all data frames.

The C-step enforces consistency between the updated mean photon numbers, W'_{ij}, by mapping them back to reciprocal space to form a new 3D intensity model, $W'(\mathbf{q})$. Recall that W'_{ij} is the intensity value sampled from $W'(\mathbf{q})$ by pixel i at particle orientation j. Therefore, the mapping is given by the interpolation

$$W'(\mathbf{p}) = \frac{\sum_i \sum_j f(\mathbf{p} - \mathbf{R}_j \cdot \mathbf{q}_i) W'_{ij}}{\sum_i \sum_j f(\mathbf{p} - \mathbf{R}_j \cdot \mathbf{q}_i)}. \tag{2.47}$$

Since each voxel of $W'(\mathbf{q})$ is sampled by multiple pairs of $(\mathbf{q}_i, \mathbf{R}_j)$, the C-step improves the signal-to-noise ratio (SNR) of the voxel values by averaging over W'_{ij}. The construction of $W'(\mathbf{q})$ completes an iteration of the EMC algorithm, and the iterations continue until the 3D intensity model converges: $W \simeq W'$.

2.3.2 Different likelihood models

In Section 2.3.1, we assumed that the signal measured by each pixel follows Poisson statistics, and derived the explicit expression of $R_{jk}(W)$, the discrete representation of $p(K|\Omega, W)$, in Equation (2.42). In fact, the EMC algorithm is flexible enough to accommodate different experimental conditions by changing the definitions of $R_{jk}(W)$. This section gives a short review of some of these experimental conditions and the appropriate likelihood models used by the EMC algorithm in data analysis.

Fluctuating fluence

In many real-world applications, the signals measured in each diffraction pattern, k, fluctuate by an overall scale factor, ϕ_k, for example, the shot-to-shot fluence fluctuation at XFELs, or the volumes of different crystals in serial crystallography. Assume that p_i is the product of the polarization factor and solid angle of pixel i. Given sample orientation j, the photon count, K_{ik}, is the Poisson sample of the mean photon number

$$\widetilde{W}_{ijk} = p_i \phi_k W_{ij}. \tag{2.48}$$

Accordingly, the expected log-likelihood function can be rewritten as

$$Q(W', \phi') = \sum_j \sum_k P_{jk}(W, \phi_k) \left(\sum_i K_{ik} \log(p_i \phi'_k W'_{ij}) - (p_i \phi'_k W'_{ij}) \right), \tag{2.49}$$

where

$$P_{jk}(W, \phi_k) = \frac{w_j \prod_i \widetilde{W}_{ijk}^{K_{ik}} \exp(-\widetilde{W}_{ijk})}{\sum_{j'} w_{j'} \prod_i \widetilde{W}_{ij'k}^{K_{ik}} \exp(-\widetilde{W}_{ij'k})}. \tag{2.50}$$

When the values of ϕ_k can be estimated heuristically, the tomogram values are updated in the M-step simply by maximizing $Q(W', \phi)$ with respect to W'_{ij}:

$$W_{ij} \to W'_{ij} = \frac{\sum_k P_{jk}(W, \phi_k) K_{ik}/p_i}{\sum_k P_{jk}(W, \phi_k)\phi_k}. \tag{2.51}$$

In the C-step, the tomograms, W'_{ij}, are weighted by $\sum_k P_{jk}(W, \phi_k)\phi_k$ to reflect the frequency of orientation j populated by the data frames with weight corresponding to the signal strength of the frame:

$$W'(\mathbf{p}) = \frac{\sum_i \sum_j f(\mathbf{p} - \mathbf{R}_j \cdot \mathbf{q}_i) \left(\sum_k P_{jk}(W, \phi_k)\phi_k \right) W'_{ij}}{\sum_i \sum_j f(\mathbf{p} - \mathbf{R}_j \cdot \mathbf{q}_i) \left(\sum_k P_{jk}(W, \phi_k)\phi_k \right)}. \tag{2.52}$$

Often the values of ϕ_k have to be reconstructed along with the 3D intensity model, W. However, simultaneous updates for W' and ϕ' are nontrivial because they appear as products in $Q(W', \phi')$. We instead update the models by maximizing $Q(W', \phi')$ with one or the other parameter, W' or ϕ', held fixed in each EMC iteration [45], which gives the update rules

$$W_{ij} \to W'_{ij} = \frac{\sum_k P_{jk}(W, \phi_k) K_{ik}/p_i}{\sum_k P_{jk}(W, \phi_k)\phi_k} \tag{2.53}$$

$$\phi_k \to \phi'_k = \frac{\sum_j P_{jk}(W, \phi_k) \sum_i K_{ik}}{\sum_j P_{jk}(W, \phi_k) \sum_i p_i W_{ij}}. \tag{2.54}$$

If the updated model is W' in an iteration, we implement the C-step using Equation (2.52); otherwise, the current model, ϕ, is replaced by ϕ' to start the next iteration.

Extraneous background

A more realistic signal model considers each measured diffraction pattern as the Poisson sample of the incoherent sum of diffuse background and the sample diffraction at some

orientation. Specifically, the mean photon number, $\widetilde{W}_{ijk}$, is modeled as

$$\widetilde{W}_{ijk} = b_{ik} + p_i \phi_k W_{ij}, \tag{2.55}$$

where b_{ik} is the background estimate at pixel i in data frame k. With this signal model, the expected log-likelihhod function, $Q(W', \phi')$, is given by

$$Q(W', \phi') = \sum_j \sum_k P_{jk}(W, \phi_k) \left(\sum_i K_{ik} \log(b_{ik} + p_i \phi'_k W'_{ij}) - (b_{ik} + p_i \phi'_k W'_{ij}) \right), \tag{2.56}$$

where

$$P_{jk}(W, \phi_k) = \frac{w_j \prod_i \widetilde{W}_{ijk}^{K_{ik}} \exp(-\widetilde{W}_{ijk})}{\sum_{j'} w_{j'} \prod_i \widetilde{W}_{ij'k}^{K_{ik}} \exp(-\widetilde{W}_{ij'k})}. \tag{2.57}$$

As discussed above, we update the models in the M-step by maximizing $Q(W', \phi')$ with one or the other parameter, W' or ϕ', held fixed in each EMC iteration. This alternating update rule converts the original problem into two sets of minimizations

$$W'_{ij} = \arg\min_{W'_{ij}} \sum_k P_{jk}(W, \phi_k) \left[(b_{ik} + p_i \phi_k W'_{ij}) - K_{ik} \log(b_{ik} + p_i \phi_k W'_{ij}) \right] \tag{2.58}$$

$$\phi'_k = \arg\min_{\phi'_k} \sum_{ij} P_{jk}(W, \phi_k) \left[(b_{ik} + p_i \phi'_k W_{ij}) - K_{ik} \log(b_{ik} + p_i \phi'_k W_{ij}) \right]. \tag{2.59}$$

When the quantities b_{ik}, p_i, ϕ_k and W_{ij} are all positive, which is strictly enforced for b_{ik} and p_i, the functions to be minimized in Equations (2.58) and (2.59) are convex, and the minima can be readily found by a line search, i.e., a simple numerical algorithm to locate minima in 1D [60]. We also impose a non-negativity constraint on ϕ'_k when solving Equation (2.59). On the other hand, W'_{ij} are allowed to be negative when solving Equation (2.58) as a result of noise, and the summation in Equation (2.59) only sums over the pairs (i, j) where the values of W_{ij} are non-negative. In the iterations that update W', the new 3D intensity model, $W'(\mathbf{p})$, is constructed in the C-step using Equation (2.52); otherwise, the current model, ϕ, is replaced by ϕ' to start the next iteration.

Gaussian noise model

Gaussian noise models were adopted in the early applications of the EMC algorithm in XFEL experiments [16, 45], possibly because of the detector readout noise or the uncertainties introduced by background subtraction. Similar to the original EMC paper, the corrections of the polarization factors and solid angles were neglected. Given the mean photon numbers, W_{ij}, the likelihood for data frame k to measure the photon counts, K_{ik}, is based on a Gaussian model:

$$R_{jk}(W, \phi_k) \propto \prod_i \exp(-\frac{(K_{ik}/\phi_k - W_{ij})^2}{2\sigma_{ij}^2}), \tag{2.60}$$

where σ_{ij} denotes the standard deviation of each Gaussian distribution, which was estimated heuristically in Ref. [16] and [45]. The expected log-likelihood function is defined by (apart from an irrelevant constant)[2]

$$Q(W', \phi') = \sum_j \sum_k P_{jk}(W, \phi_k) \left(- \sum_i \frac{(K_{ik}/\phi'_k - W'_{ij})^2}{2\sigma_{ij}^2} \right), \tag{2.61}$$

where

$$P_{jk}(W, \phi_k) = \frac{w_j R_{jk}(W, \phi_k)}{\sum_{j'} w_{j'} R_{j'k}(W, \phi_k)}. \tag{2.62}$$

Maximizing $Q(W', \phi')$ with one or the other parameter, W' or ϕ', held fixed in each EMC iteration, we obtain the alternating update rules:

$$W_{ij} \to W'_{ij} = \frac{\sum_k P_{jk}(W, \phi_k) K_{ik}/\phi_k}{\sum_k P_{jk}(W, \phi_k)} \tag{2.63}$$

$$\phi_k \to \phi'_k = \frac{\sum_j P_{jk}(W, \phi_k) \sum_i K_{ik}^2/\sigma_{ij}^2}{\sum_j P_{jk}(W, \phi_k) \sum_i K_{ik} W_{ij}/\sigma_{ij}^2}. \tag{2.64}$$

[2] In Ref. [45], the function $Q(W', \phi')$ was instead defined as

$$Q(W', \phi') = \sum_j \sum_k w_j R_{jk}(W, \phi_k) \log R_{jk}(W', \phi'_k),$$

with σ_{ij} replaced by a global parameter, σ, in Equation (2.60). Under the approximation that each data frame, k, has similar mean likelihood value, $\sum_j w_j R_{jk}(W, \phi_k)$, the update rules in Equations (2.63) and (2.64) reduce to those generated by maximizing $Q(W', \phi')$ in this definition.

In the iterations that update W', the C-step is implemented by Equation (2.52); otherwise, the current model, ϕ, is replaced by ϕ' to start the next iteration.

2.3.3 Local update scheme

The most time-intensive part of the EMC algorithm is the calculation of the conditional probabilities, $P_{jk}(W, \phi)$, which has the number of operations proportional to the number of data frames, M_{data}, the number of rotation samples, M_{rot}, and the number of pixels, M_{pix}. This makes the reconstruction of high-resolution features especially challenging due to the scaling

$$M_{\text{rot}} M_{\text{pix}} \propto q_{\text{max}}^5, \tag{2.65}$$

where q_{max} denotes the maximum spatial frequency magnitude. In this section, we describe an update scheme that exploits a special property of the EMC algorithm to speed up the reconstruction at high resolution [41].

Before elaborating on the details, we first review how an EMC reconstruction converges in qualitative terms. Since the diffraction signal strength in general decays with the increase of the spatial frequency magnitude, q, the features at low-q values are first reconstructed. These low-q features give each data frame a strong preference for certain orientations, and the 3D intensity model, W, is refined about these probable orientations to resolve features at higher resolution. With improved SNR in W, the convergence gradually proceeds from low-q to high-q values. This observation shows that the intensity reconstruction has the property of locality in orientations: each data frame, k, has high probabilities, P_{jk}, only at a handful of orientations favored by the low-q features; on the other hand, the other orientations with negligible probabilities hardly contribute to the refinement of W. Therefore, the computation time can be significantly reduced by

restricting the search to the vicinity of the probable orientations on a per frame basis.

The computing scheme that we call the local update scheme takes advantage of the locality in orientations to speed up the convergence of the EMC reconstruction. Given a converged 3D intensity model, W, with a coarse rotation sampling of order n_c (labeled by j_c) and the conditional probabilities, $P_{j_c k}(W, \phi_k)$, for each data frame, k, we first represent the probable orientation list by a binary matrix

$$B_{j_c k} = \begin{cases} 1, & \text{if } P_{j_c k}(W, \phi_k) > \epsilon_p \\ 0, & \text{otherwise} \end{cases}, \tag{2.66}$$

where ϵ_p is a predefined threshold. Our aim is to refine W using a fine rotation sampling of order n_f (labeled by j_f) without calculating all the elements of $P_{j_f k}(W, \phi_k)$. For each coarse rotation sample, j_c, we define its neighborhood as the subset of rotation space that is closer to j_c than any other samples, and assign the fine rotation samples, j_f, that lie in this subset as the neighbors of j_c. This mapping is stored as a matrix

$$C_{j_c j_f} = \begin{cases} 1, & \text{if } j_f \text{ is a neighbor of } j_c \\ 0, & \text{otherwise} \end{cases}. \tag{2.67}$$

Subsequently, we refine W in the usual way of the EMC algorithm, with the exception that only the entries of $P_{j_f k}(W, \phi_k)$ that satisfy the conditions $B_{j_c k} = 1$ and $C_{j_c j_f} = 1$ are calculated while the others are set to zero. This requirement restricts the calculation of $P_{j_f k}(W, \phi_k)$ to the neighbors of the probable coarse rotation samples of each data frame.

Restricting the search in orientations significantly speeds up the EMC reconstruction. Assume that each data frame on average has N_p coarse rotation samples with non-negligible probabilities. Since the sampling density of rotations is proportional to the cube of the order, n, the local update scheme on average calculates $N_p n_f^3 / n_c^3$ entries of $P_{j_f k}(W, \phi_k)$ per frame. This speed-up corresponds to a factor of n_c^3 / N_p. Moreover, the matrices $B_{j_c k}$ and $C_{j_c j_f}$ are both sparse, and barely add any burden to the memory usage.

The idea of the local update scheme is similar to the sparse update scheme proposed by Neal & Hinton [53], which speeds up the expectation-maximization algorithm by leaving out the improbable values of the searched parameters in most of the iterations and only updating them at a much lower rate. The only difference between the two schemes is the specific property of locality in orientations in our intensity reconstruction application, which allows us to search in a finer grid about the probable coarse rotation samples to refine W at high resolution. Nonetheless, we need to stress that the only reason to adopt the local update scheme is to speed up the reconstruction at high resolution. The likelihood function maximized with the local update restriction cannot exceed its counterpart when the whole rotation group is explored.

2.3.4 Memory-efficient parallel implementation

The memory usage of the EMC algorithm is dominated by the conditional probabilities, $P_{jk}(W, \phi_k)$, and the tomogram values, W_{ij} and W'_{ij}, which have sizes of $M_{\text{rot}} \times M_{\text{data}}$ and $M_{\text{pix}} \times M_{\text{rot}}$, respectively. Since $M_{\text{rot}} \propto q_{\text{max}}^3$ and $M_{\text{pix}} \propto q_{\text{max}}^2$, the required memory rapidly becomes prohibitive even with modest angular resolution [3]. In this section, we introduce a parallel implementation of the EMC algorithm that allows high-resolution reconstructions with reasonable memory usage.

We first notice that each data frame only has non-negligible probabilities at a few orientations, unless the signal level is as weak as just several photons per frame. Therefore, the entries of $P_{jk}(W, \phi_k)$ can be stored as a sparse matrix to save memory. In our implementation, we distribute blocks of data frames (ranges in k index) to different processors, each of which holds the same copies of models, W and ϕ, and the algorithm strides through the M_{rot} rotations in steps of size M_{step} to calculate W_{ij} and $R_{jk}(W, \phi_k)$.

Each processor dynamically updates the value of $\max_j w_j R_{jk}(W, \phi_k)$ for each data frame when walking through all the orientations. From the inequality

$$P_{jk}(W, \phi_k) = \frac{w_j R_{jk}(W, \phi_k)}{\sum_{j'} w_{j'} R_{j'k}(W, \phi_k)} \leq \frac{w_j R_{jk}(W, \phi_k)}{\max_{j'} w_{j'} R_{j'k}(W, \phi_k)}, \tag{2.68}$$

the entries of $w_j R_{jk}(W, \phi_k)$ are saved only when

$$\frac{w_j R_{jk}(W, \phi_k)}{\max_{j'} w_{j'} R_{j'k}(W, \phi_k)} > \epsilon_p, \tag{2.69}$$

where ϵ_p is a predefined threshold. This condition is checked for all the saved entries every time the value $\max_j w_j R_{jk}(W, \phi_k)$ is updated. After going through all the rotation samples, the algorithm calculates the significant values of $P_{jk}(W, \phi_k)$ by normalizing the saved entries of $w_j R_{jk}(W, \phi_k)$ over orientations. Subsequently, we update W'_{ij} also in steps of size M_{step} over all the orientations. The values of W'_{ij} are mapped back to the updated 3D intensity model, W', after each step.

In our implementation the memory usage is dominated by W_{ij} (W'_{ij}), and scales as $N_{\text{proc}} \times M_{\text{pix}} \times M_{\text{step}}$, where N_{proc} denotes the number of processors. This new memory scaling is only proportional to q^2_{max}, and can in practice limit the total memory usage to only tens to hundreds of gigabytes (GB) even with very high angular resolution.

CHAPTER 3

SINGLE PARTICLE IMAGING

A desirable goal in structural biology using XFEL facilities is to image the 3D structure of biological macromolecules in near-physiological conditions. In SPI experiments, diffraction patterns are collected from many reasonably identical copies of a bioparticle, delivered in random orientations into the pulsed X-ray beam. The femtoseconds long pulse width enables the scattering process to outrun the structural destruction by the intense pulses [54]. The 3D structure of the bioparticle is then solved by phasing the 3D intensity volume assembled from the unoriented diffraction patterns in reciprocal space. Although still in the development stage, successful applications of SPI at sub-nanometer resolution will offer an unparalleled tool to probe the dynamics of biological macromolecules [68]. In order to resolve the technical problems that challenge SPI, an international collaboration called the SPI Initiative formed in 2015 [1]. The collaboration has carried out a few R&D experiments at the Linac Coherent Light Source (LCLS), and made considerable progress in experimental technology. Here we describe our contribution from the data analysis side.

3.1 Sample selection

The goal of the first few R&D experiments was to optimize the experimental conditions in SPI. The ideal test samples that help achieve this goal should at least have the following characteristics:

- Available in large quantities: In these early experiments, we should expect the mean hit rate for a particle to be intercepted by an X-ray pulse to be just a few percent or lower, so a large quantity of sample is needed.

- Reproducible structure to avoid the complication of sample heterogeneity.
- Known structure to validate the structure solutions.
- Ease of orientation reconstruction.

By focusing on the last point, we discuss our evaluation of the proposed nine different bioparticles. In particular, we ranked the hardness to orient the diffraction patterns of each bioparticle through computer simulation.

3.1.1 Diffraction pattern simulation

The proposed nine bioparticles for the first few SPI experiments were:

1. CalS11 methyltransferase fusion protein (PDB entry: 3TOS)
2. KLH1 di-decamer (PDB entry: 4BED [28])
3. Yeast RNA Polymerase II (PDB entry: 1WCM [2])
4. MS2 phage empty capsid (PDB entry: 1ZDI [74])
5. Tomato bushy stunt virus (TBSV) (PDB entry: 2TBV [34])
6. Four-layer tobacco mosaic virus (TMV) (PDB entry: 1EI7 [8])
7. DNA origami (PDB entry: 4V5X [6])
8. Coliphage PR772 virus (no available structure)
9. Rice dwarf virus (RDV) (PDB entry: 1UF2 [51])

These particles have sizes ranging from 10 to 70 nm, and we would like to simulate the diffraction patterns from the atomic coordinates in the PDB files. Consider an SPI experiment with X-ray wavelength λ, sample size L and sample-to-detector distance D.

The diffraction patterns are collected by a pixelated detector of squared pixel size d. We define R as the distance (in pixels) from the beam incidence point to the edge of the detector, rounded to the nearest integer, and neglect the pixels outside this radius. The maximum spatial frequency magnitude measured by the detector is then given by

$$q_{\max} = \frac{2\pi}{\lambda} \cdot 2 \sin \frac{\theta}{2}, \tag{3.1}$$

where $\theta = \tan^{-1}(Rd/D)$ is the maximum scattering angle. The pixels within radius R_{stop} are blocked to protect the detector from direct beam illumination.

From Shannon's sampling theorem [65], a 1D band-limited function $\hat{f}(q)$, where $\hat{f}(q) = 0$ when $|q| > q_{\max}$, can be fully represented by its inverse Fourier transform, $f(x)$, sampled at points x_n:

$$\hat{f}(q) = \sqrt{\frac{\pi}{2}} \frac{1}{q_{\max}} \sum_{n=-\infty}^{\infty} f(x_n)\, e^{-iqx_n}, \tag{3.2}$$

where $x_n = n\Delta x = n(\pi/q_{\max})$. In our SPI simulation, the Fourier components are band-limited by $|\mathbf{q}| < q_{\max}$ because of the finite detector size, which indicates that we have to sample the 3D contrasts in real space at a rate of

$$\Delta x = \frac{\pi}{q_{\max}}. \tag{3.3}$$

This value is also called the half-period resolution.

The 3D contrasts were constructed as follows: After binning the coordinates of the non-hydrogen atoms in a PDB file on a cubic grid of voxel size 2 Å, a discrete Fourier transform was applied to the grid and truncated at the size $2r + 1$, where each atom was weighted by its atomic number and r was given by $(L/\Delta x - 1)/2$ rounded up to the nearest integer. The truncated Fourier transform was then multiplied by a low-pass Gaussian filter

$$G(\mathbf{q}) = \exp(-2.3\, |\mathbf{q}|^2 / q_{\max}^2), \tag{3.4}$$

where $|\mathbf{q}|$ was calculated by the distance (in voxels) from the central voxel multiplied by $q_{\max}/r$. This filter models the decay of Fourier magnitudes due to the blurring of atomic positions over the X-ray pulse duration [46]. The result was inverse Fourier transformed to give the 3D contrast in real space, supported on a cubic grid of length $2r + 1$.

The diffraction patterns were simulated by the Poisson samples of the mean photon numbers, $p_i W_{ij} \Delta\Omega_i$, where p_i and $\Delta\Omega_i$ denote the polarization and solid angle corrections for pixel i, respectively. The incident X-ray was assumed to be horizontally linear polarized. The number, W_{ij}, is given by the interpolation in Equation (2.40), where j indexes the bioparticle orientations and $W(\mathbf{q})$ is the time-integrated intensity defined in Equation (2.17). After embedding the 3D contrasts constructed above on a larger cubic grid of size $2R + 1$, the squared Fourier magnitudes, $|\hat{\rho}(\mathbf{q})|^2$, were computed. Using the experimental parameters given in Table 3.1, we generated diffraction patterns with $q_{\max} = 0.4$ Å^{-1}, which corresponds to $\Delta x \approx 8$ Å.

Photon energy (keV)	6
Incident photon density, $J_{\text{inc}}\Delta t$ (photons $\cdot$ μm^{-2} $\cdot$ pulse^{-1})	10^{13}
Detector distance, D (mm)	417
Detector radius, R (pixel)	500
Beamstop radius, R_{stop} (pixel)	20
Pixel size, d (μm)	110

Table 3.1: Parameters for the SPI simulations.

3.1.2 SNR of speckles

The noise in SPI mainly consists of two parts — the statistical noise due to Poisson statistics and the systematic noise due to background scattering. We first consider two extreme cases of background noise. In one extreme, the background is strong and catches up with the particle signal at some spatial frequency magnitude, $q_{\max}$. To achieve

the highest possible resolution, the sample that scatters the strongest signal should be selected, which favors the largest particle in our list (see Section 3.1.1), RDV.

In the other extreme, the background can be reliably suppressed so that it does not compete with the particle signal. This can be done by placing an aperture downstream of the sample to block the parasitic scattering from the beamline optics [78]. If the background stays well below the particle signal, the achievable resolution will be determined by the number of particle diffraction patterns collected in the experiment. In this section, we will calculate the SNR of the proposed samples in this hit-rate limited regime.

After averaging over particle orientations, the number density of photons, $n(q)$, scattered by spatial frequency magnitude, q, into reciprocal space volume element, d^3q, scales with the particle volume, V, and a fall-off function, $s(q)$:

$$n(q) \propto s(q)V. \tag{3.5}$$

Here we assume that the incident beam fluence is fixed. For non-crystalline samples, the Fourier intensities consist of many small smooth regions known as speckles. The speckle volume is roughly homogeneous and inversely proportional to the sample volume: $\tilde{V} \propto 1/V$. Suppose we succeeded in collecting H particle diffraction patterns and that they were correctly classified and combined to form a 3D intensity map. The number of photons that contribute to a speckle at q is given by

$$N(q) \propto n(q)\tilde{V}H \propto s(q)H, \tag{3.6}$$

and we can readily obtain the q-dependence of the SNR from Poisson statistics:

$$SNR(q) \propto \sqrt{N(q)} \propto \sqrt{s(q)H}. \tag{3.7}$$

The resolution cutoff, $q_{\max}$, is therefore determined by the equation

$$SNR(q_{\max}) \propto \sqrt{s(q_{\max})H}, \tag{3.8}$$

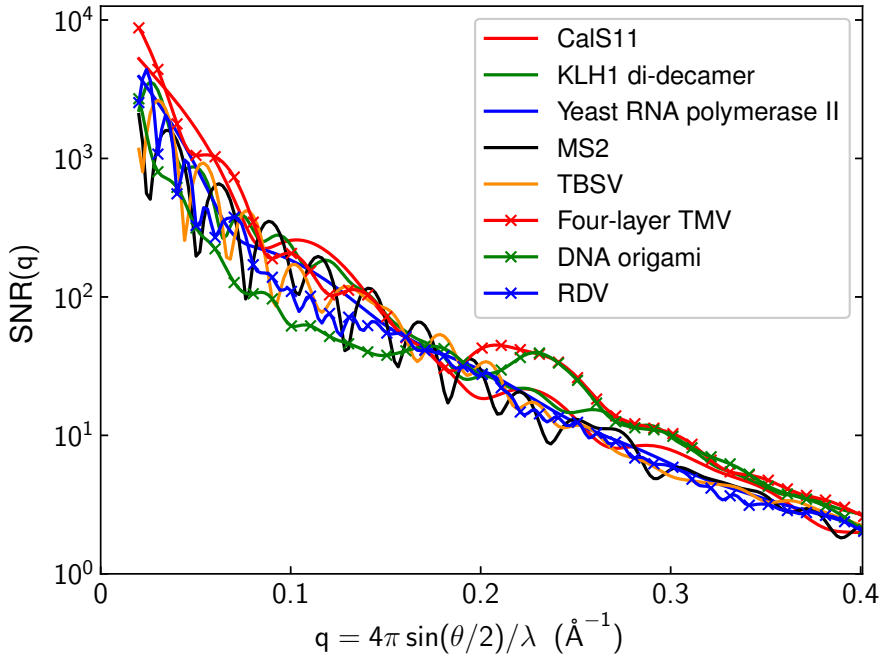

Figure 3.1: Resolution dependent SNR of speckles for the proposed SPI samples. Each dataset has 10^5 hits and is assumed to be correctly merged in reciprocal space. By setting SNR = 20 as a general criterion, the resolution limit is $\Delta x = 14$ Å ($q_{\max} = 0.22$ Å^{-1}).

when the SNR falls below a threshold value. Because the fall-off function, $s(q)$, is characteristic of the material (biomolecule in our case) and independent of particle size, we expect the achievable resolution in the hit-rate limited regime to only depend on the number of particle hits, H.

Using the experimental parameters in Table 3.1, we simulated $H = 10^5$ randomly oriented diffraction patterns of each proposed sample except for PR772 virus to calculate the number density of photons, $n(q)$. We used $\tilde{V} = (2\pi)^3/V$ for the speckle volume, and defined the resolution by requiring the SNR of the outermost speckles,

$$SNR(q) = \sqrt{n(q)\tilde{V}H}, \tag{3.9}$$

to be above some lower limit. The results are shown in Figure 3.1. As argued above, the SNR of speckles is not a discriminative criterion for sample selection. By setting the limit as SNR = 20, the achievable half-period resolution is $\Delta x = \pi/q_{\max} \approx 14$ Å.

3.1.3 Hardness of orientation reconstruction

In the discussion above we did not consider the process of merging diffraction patterns, but only asked that the final 3D intensity map should have sufficient SNR for convergence to be possible at the highest resolution. This overlooks the very daunting problem that initially, before we have anything resembling speckles, we have very poor information for assigning even tentative orientations to the diffraction patterns. For the purpose of sample selection, we would like to know if the diffraction patterns of some samples merge more easily than others. Before answering this question, it is helpful to first understand how the diffraction patterns are merged in qualitative terms.

The diffraction signal drops rapidly with the increase of q, so the small-q speckles, which generally have higher SNR, are reconstructed first. Due to their larger angular sizes, the early-stage orientation assignment tends to have larger errors. With the improved SNR in the low-q speckles, the orientations become more accurate and the refinement of the 3D intensity map gradually proceeds to reconstruct the high-q speckles.

The qualitative description of orientation reconstruction suggests two factors that determine the hardness of merging diffraction patterns. Clearly the angular variation of the intensity is one of them: greater variation translates to greater angular information. The other factor is the signal strength of the diffraction patterns, because the orientational information is degraded by the Poisson noise of photon detection. The combined effect of these factors is captured by a form of mutual information — a measure of the orientational information gained, on average, by the detection of photons.

Consider photons detected in resolution shell q. Let w be the mean photon number measured by a pixel in this resolution shell for some particle orientation. The angular average, $\langle w \rangle$, gives the average photon number detected per pixel in this resolution shell.

Using these two quantities, we can construct the mutual information (derived in the end of this section)

$$\Omega = \langle w \rangle \left\langle f\left(\frac{w}{\langle w \rangle}\right)\right\rangle, \tag{3.10}$$

where $f(x) = x \log x - x + 1$. The quantity Ω represents the information gained, per pixel, on the angular distribution of intensity in resolution shell q by measuring photons. Dividing Ω by $\log 2$ expresses this information in units of bits. If there are $m(q)dq$ detector pixels between resolution q and $q + dq$, then on average we obtain $\Omega(q)m(q)dq/\log 2$ bits of angular information from the photons detected in this shell.

We can get an intuitive sense of the expression for Ω by a simple approximation. Suppose that the angular variation of w is small, as in the case of an icosahedral virus at small q. In that case $x = w/\langle w \rangle \approx 1$ and we can expand $f(x)$ for x near 1:

$$f(x) \approx \frac{1}{2}(x-1)^2. \tag{3.11}$$

Using this approximation, we obtain

$$\Omega = \frac{1}{2}\langle w \rangle^{-1} (w - \langle w \rangle)^2, \tag{3.12}$$

which represents the angular variance of the intensity with an overall scale that goes as the mean intensity, $\langle w \rangle$.

In Figure 3.2 we have plotted $\Omega(q)$ as a function of resolution for all the proposed samples except for PR772 virus. Perhaps contrary to expectations, one of the best samples by this metric is an icosahedral virus — RDV. The overall scale of the intensity more than makes up for the low angular variation. What appears as another promising candidate is KLH1 di-decamer, in this case thanks to its more pronounced angular variation. An argument can be made for TBSV or DNA origami, but both of these are deficient in information relative to the other two at small q, where data merging begins.

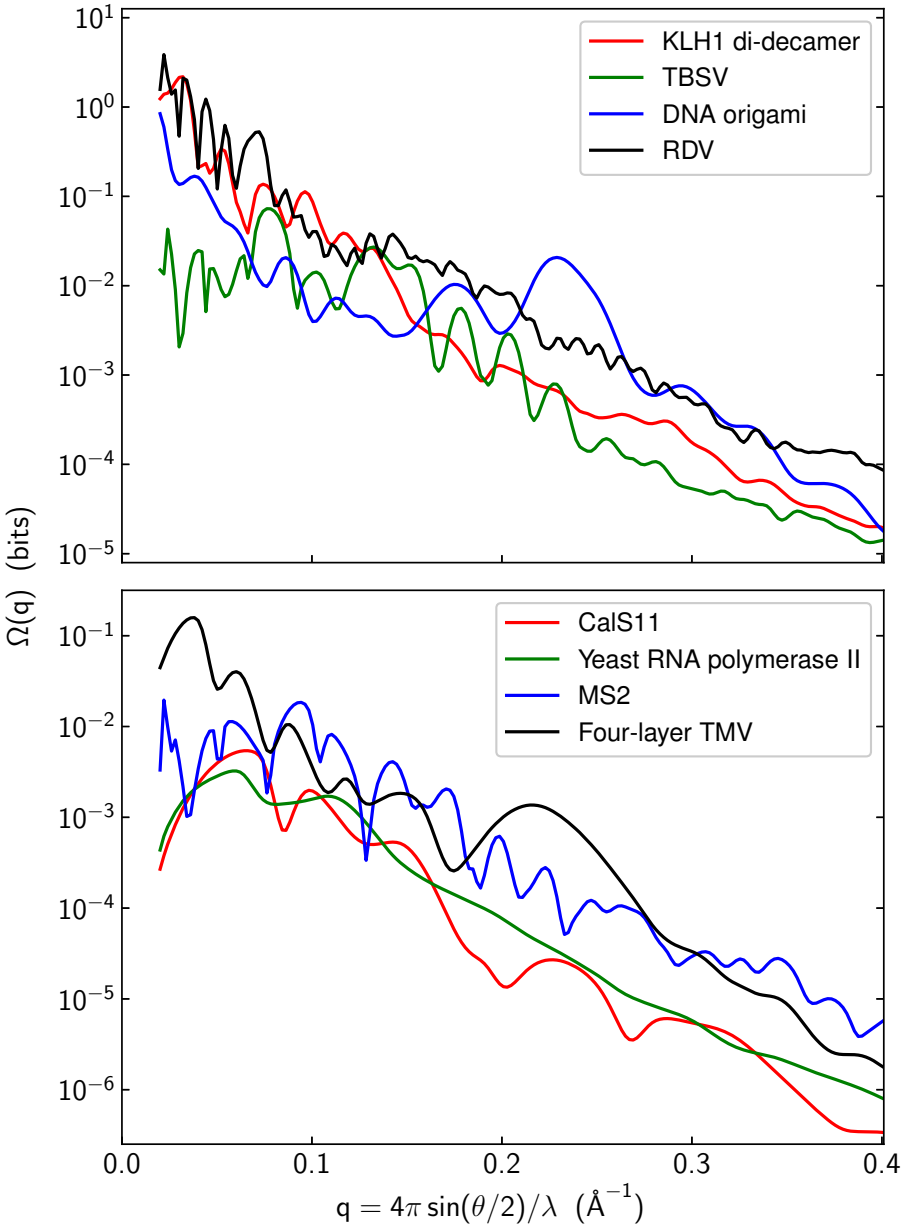

Figure 3.2: Mutual information measure of hardness of orientation reconstruction as a function of resolution for the proposed SPI samples. The orientation information gain per diffraction pattern for the samples in the top figure is about one order of magnitude higher than that for the samples in the bottom figure.

A related metric is the total orientational information one measures over all the resolution shells up to the maximum resolution q:

$$\Omega_{\text{total}}(q) = \int_0^q \Omega(q')m(q')dq', \tag{3.13}$$

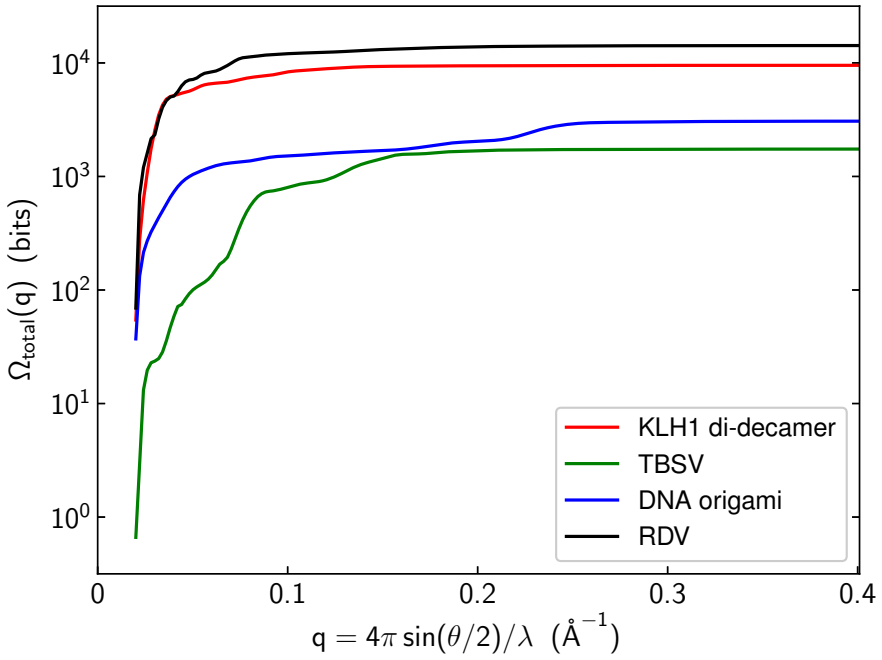

Figure 3.3: Integrated orientational information over resolution in a diffraction pattern for the four top candidates by the Ω metric.

where $m(q)dq$ is the number of detector pixels between resolutions q and $q + dq$. This quantity is plotted in Figure 3.3 for the four top candidates by the Ω metric. We can see that RDV and KLH1 are again at the top, and most of the information comes from very small q. To put the numbers on the vertical axis in perspective, we note that resolving a single angle to one degree requires about 8 bits of information, so orienting a single frame (three Euler angles) requires on the order of 25 bits. Figure 3.3 shows that this quantity of information would be absent if the beamstop eliminated the sharp initial rise of Ω_{total} at small q.

To summarize, we advocated RDV and KLH1 from a theoretical viewpoint. Their dominance in the metrics of Ω and Ω_{total} shows that the diffraction patterns of these two particles carry higher information content — a sign for easier orientation reconstruction.

Derivation of mutual information Ω

Consider the joint probability of two random variables: K, the photon count measured by a particular pixel, and θ, the angular position[1] in the resolution shell that the pixel is measuring for some random orientation of the particle. The conditional probability, of K being measured given a particular θ, is

$$p(K|\theta) = \begin{cases} w(\theta), & K = 1 \\ 1 - w(\theta), & K = 0 \end{cases}, \tag{3.14}$$

where $w(\theta)$ is the mean photon number measured by the pixel. We are taking the $w(\theta) \ll 1$ limit of the Poisson distribution in these formulae, a valid approximation for all the samples under consideration. The joint distribution, $p(K, \theta) = p(K|\theta)p(\theta)$, is proportional to the conditional distribution above, since the orientation distribution of the particle is assumed to be uniform. The marginal distribution of the photon count is formed by integrating the joint distribution over θ:

$$p(K) = \begin{cases} \langle w \rangle, & K = 1 \\ \langle 1 - w \rangle, & K = 0 \end{cases}, \tag{3.15}$$

where the angle brackets denote a uniform average over θ.

From the conditional distribution, we obtain the conditional entropy of photon counts by computing the entropy of K given some θ and averaging over θ:

$$\begin{aligned} H(K|\theta) &= \langle -p(1|\theta) \log p(1|\theta) - p(0|\theta) \log p(0|\theta) \rangle \\ &= \langle -w \log w - (1 - w) \log(1 - w) \rangle \\ &\approx \langle w - w \log w \rangle, \end{aligned} \tag{3.16}$$

where in the last line only the leading terms are kept in the limit $w(\theta) \ll 1$. The entropy

[1] Here θ is a generic angular position index for the shell, not the polar angle.

of the photon counts, without conditions, is obtained from the marginal distribution:

$$\begin{aligned} H(K) &= -p(1)\log p(1) - p(0)\log p(0) \\ &= -\langle w\rangle \log\langle w\rangle - \langle 1-w\rangle \log\langle 1-w\rangle \\ &\approx \langle w\rangle - \langle w\rangle \log\langle w\rangle\,. \end{aligned} \tag{3.17}$$

The mutual information associated with our pair of random variables, K and θ, is defined as the difference of entropies:

$$\Omega = I(K,\theta) = H(K) - H(K|\theta) \tag{3.18}$$

$$= H(\theta) - H(\theta|K). \tag{3.19}$$

The second form is easiest to interpret in our context. The first term $H(\theta)$ represents the number of bits of information[2] associated with our angular resolution — the maximum information we could hope to obtain through the measurement at one pixel. In the $w(\theta) \ll 1$ limit, however, there are only two outcomes of the measurement ($K = 0$ or $K = 1$) and consequently there is a large entropy in the possible angles that could have produced the measurement. The number of bits (entropy) associated with this uncertainty, $H(\theta|K)$, gets subtracted from the number of bits in our angular resolution, $H(\theta)$, to yield the actual number of bits of information gained by the measurement.

Substituting Equations (3.16) and (3.17) into Equation (3.18), we obtain

$$\begin{aligned} \Omega &= \langle w\log w\rangle - \langle w\rangle\log\langle w\rangle \\ &= \left\langle w\log\frac{w}{\langle w\rangle}\right\rangle \\ &= \langle w\rangle\left\langle \frac{w}{\langle w\rangle}\log\frac{w}{\langle w\rangle}\right\rangle \\ &= \langle w\rangle\left\langle \frac{w}{\langle w\rangle}\log\frac{w}{\langle w\rangle} - \frac{w}{\langle w\rangle} + 1\right\rangle \end{aligned}$$

[2]To get units of bits we need to divide Ω by $\log 2$.

$$= \langle w \rangle \left\langle f\left(\frac{w}{\langle w \rangle}\right)\right\rangle . \tag{3.20}$$

3.2 Data analysis

Here we describe the results of our analysis on the data collected in two SPI experiments. In the analysis of the first dataset, we developed a metric to measure the agreement of the data with a known model. In the other analysis, we applied the EMC algorithm to reconstruct the 3D intensity map, from which we solved the structure by phasing.

3.2.1 Normalized surprise function

The first dataset we analyzed [49] was taken from RDV particles at the CXI beamline [44] of the LCLS in June, 2015. It was challenging to reconstruct a 3D intensity map from the data due to the limited number of single-particle hits (175 determined by Hummingbird [13] by counting photons in a region of interest on one of the detectors). The scarcity of diffraction patterns was caused by the difficulty to inject the virus particles into the 100 nm wide focus of the X-ray beam. Nevertheless, the data quality can still be examined using the known structure of RDV (PDB entry: 1UF2).

Figure 3.4 shows the diffraction patterns of one of the 175 selected single-particle hits. The diffraction patterns were collected by two detectors arranged in tandem, where the central hole of the front detector allows the scattered photons at low-q to pass through and be recorded by the back detector. The incident photons had energy 7 keV, and the detector distances were 217.4 mm and 2.4 m for the front and back detectors, respectively. Our analysis focused on the front detector, where we used data up to the

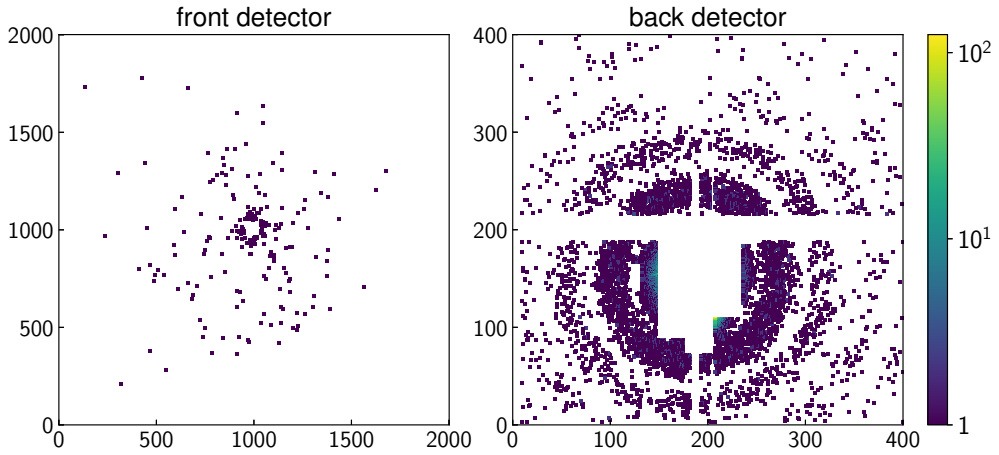

Figure 3.4: Diffraction patterns of a single-particle hit of RDV recorded by the front and back detectors. Each dot represents the photon count detected by a pixel. The front and back detectors collected about 200 and 9,000 photons in total, respectively. The regions of beamstop and gaps on the back detector are masked out.

half-period resolution of $\Delta x = 6.67$ Å, or a radius of 265 pixels. For the 70.8-nm sized RDV particles, this value corresponds to a subdivision of 107 resolution elements across the diameter of RDV.

Following the procedure described in Section 3.1.1, we simulated the 3D Fourier intensity of RDV from the PDB file, 1UF2. Assuming Poisson statistics, we define the surprise function S as the negative of the log-likelihood function

$$S(K;\Phi,\Omega_j) = -\sum_{i=1}^{N_{\text{pix}}} \log\left(\frac{n_i^{k_i} e^{-n_i}}{k_i!}\right), \tag{3.21}$$

where K denotes the dependence on data, with k_i being the measured photon count at pixel i, n_i is the mean photon number at pixel i when the fluence value is Φ and the RDV particle has orientation Ω_j. The surprise of each frame was evaluated at different orientations across several fluence parameters, and we assigned each data frame with the orientation and fluence value minimizing the surprise (maximizing the log-likelihood).

To put the minimum surprise values on an absolute scale, we further 'normalize' the

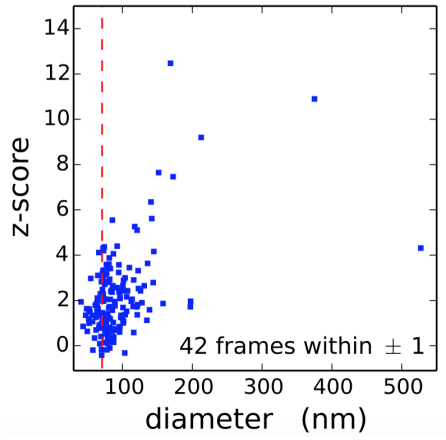

Figure 3.5: Front detector normalized surprise (z-score) versus back detector particle size fits. The dashed red line indicates the diameter (70.8 nm) of RDV. The normalized surprise function, or its z-score, measures the agreement of the data with a known model — A data frame is 'surprising' given the assumed model when the absolute value of its z-score is much greater than unity.

surprise function. We define the expectation value of the surprise function as

$$H(\Phi, \Omega_j) = -\sum_{i=1}^{N_{\text{pix}}} \sum_{k=0}^{\infty} \frac{n_i^k e^{-n_i}}{k!} \log\left(\frac{n_i^k e^{-n_i}}{k!}\right), \tag{3.22}$$

and the variance of the surprise function is given by

$$\sigma_S^2(\Phi, \Omega_j) = \sum_{i=1}^{N_{\text{pix}}} \left[\sum_{k=0}^{\infty} \frac{n_i^k e^{-n_i}}{k!} \left(\log\left(\frac{n_i^k e^{-n_i}}{k!}\right)\right)^2 - \left(\sum_{k=0}^{\infty} \frac{n_i^k e^{-n_i}}{k!} \log\left(\frac{n_i^k e^{-n_i}}{k!}\right)\right)^2 \right]. \tag{3.23}$$

It is notable that $H(\Phi, \Omega_j)$ is exactly the entropy of the photon counts when the fluence is Φ and the RDV particle has orientation Ω_j, and $H(\Phi, \Omega_j)$ and $\sigma_S(\Phi, \Omega_j)$ are independent of the data, K. The normalized surprise function, or its z-score,

$$z(K; \Phi, \Omega_j) = \frac{S(K; \Phi, \Omega_j) - H(\Phi, \Omega_j)}{\sigma_S(\Phi, \Omega_j)} \tag{3.24}$$

measures the agreement of the data with a known model — The data is inconsistent with the model when the absolute value of the z-score is much greater than unity.

The z-scores of the 175 selected frames versus particle sizes are shown in Figure 3.5. The particle sizes were determined by fitting back detector data to a homogeneous sphere

with adjustable size and a mass density of 1.381 g/cm^3. Frames with particle sizes close to the diameter of RDV generally have smaller z-scores, though some still manifest inconsistency with the model. This could be caused by the presence of a water layer on the particle surface. This model-based normalized surprise function may potentially be useful for hit-finding, especially when a model of similar structure is available and the particle is too small to produce a recognizable signal on the back detector.

3.2.2 Structure reconstruction

The second dataset we analyzed [61] was collected from PR772 virus particles at the AMO beamline [21] of the LCLS in August, 2015. The diffraction patterns were collected at photon energy of 1.6 keV by two detectors arranged in tandem. The detector distances were 100 mm and 581 mm for the front and back detectors, respectively. Due to the dysfunction of part of the front detector panels, the back detector data was used for our structure reconstruction. A total of 16,859 frames were used in the reconstruction, with one of them shown in Figure 3.6.

Due to the fluence fluctuation of XFEL sources from shot to shot, we modeled the mean photon number, $\widetilde{W}_{ijk}$, measured by pixel i in data frame k when the particle has orientation j by (see Section 2.3.2 for more details)

$$\widetilde{W}_{ijk} = p_i \phi_k W_{ij}, \tag{3.25}$$

where p_i is the product of the polarization factor and solid angle of pixel i, ϕ_k is a scale factor that accounts for the fluence fluctuation in data frame k, and W_{ij} is the tomogram value calculated from the 3D intensity model, W.

Since the icosahedral PR772 virus can be approximated as a sphere at low resolution,

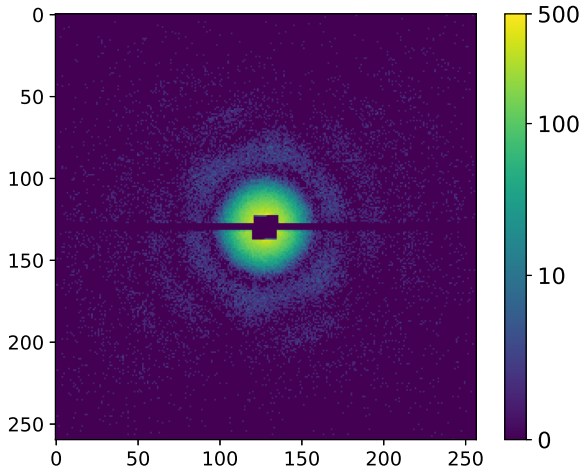

Figure 3.6: Diffraction pattern of a single-particle hit of PR772 virus collected by the back detector in units of photons. The regions of the beamstop and detector gap are masked out. The half-period resolutions are 5.8 nm at the edge and 4.2 nm in the corner of the detector.

we estimated the per-frame scale factor, ϕ_k, by

$$\phi_k = \frac{\sum_i K_{ik}/p_i}{\sum_i \sum_j w_j W_{ij}} \tag{3.26}$$

after the E-step in each iteration of the EMC reconstruction, where K_{ik} is the photon count measured by pixel i in data frame k, and w_j is the fraction of the continuous rotation group assigned to rotation sample j. Using the estimated values of ϕ_k, the tomograms were updated in the M-step by

$$W_{ij} \rightarrow W'_{ij} = \frac{\sum_k P_{jk}(W, \phi_k) K_{ik}/p_i}{\sum_k P_{jk}(W, \phi_k)\phi_k}, \tag{3.27}$$

where the conditional probabilities, $P_{jk}(W, \phi_k)$, are given by

$$P_{jk}(W, \phi_k) = \frac{w_j \prod_i \widetilde{W}_{ijk}^{K_{ik}} \exp(-\widetilde{W}_{ijk})}{\sum_{j'} w_{j'} \prod_i \widetilde{W}_{ij'k}^{K_{ik}} \exp(-\widetilde{W}_{ij'k})}. \tag{3.28}$$

The updated tomograms, W'_{ij}, were merged in the C-step using Equation (2.52) to form a new 3D intensity model, W', and then the Friedel symmetry was imposed. The EMC iterations continued until the 3D intensity model converged, whose central slices are shown in Figure 3.7.

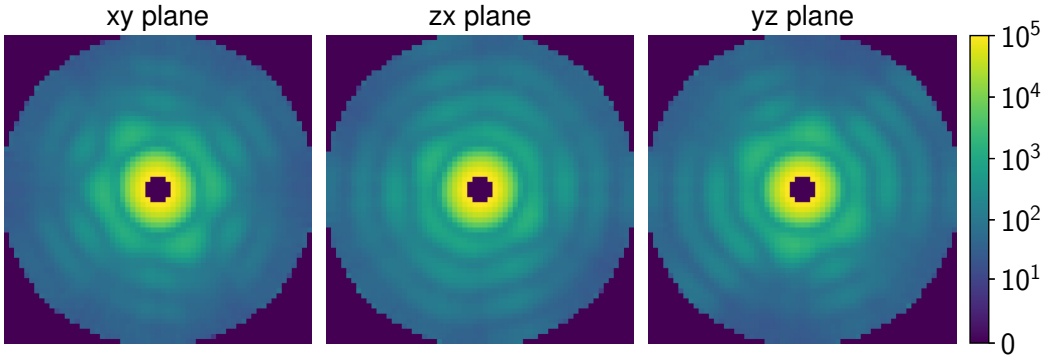

Figure 3.7: Central slices of the reconstructed 3D intensity model of PR772 virus, rendered in arbitrary units. The highest resolution cutoff corresponds to the half-period resolution of 5.9 nm.

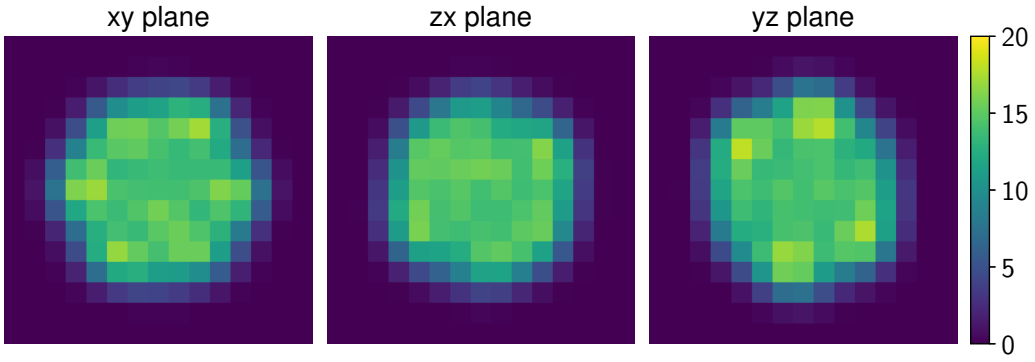

Figure 3.8: Central slices of the reconstructed real-space contrast of PR772 virus, rendered in arbitrary units, by phasing the reconstructed 3D intensity model shown in Figure 3.7. With the half-period resolution of $\Delta x = 5.9$ nm, the particle size can be estimated to be about 70 nm.

The phasing step was done using the difference map algorithm [17]. By applying an inverse FFT on the converged 3D intensity model, we obtained the particle autocorrelation, from which we estimated the particle support size. Using simple support and Fourier magnitude projections, the difference map algorithm was implemented for several thousands of iterations to reconstruct the particle contrast in real space. The central slices of the 3D contrast are rendered in Figure 3.8. With the half-period resolution of $\Delta x = 5.9$ nm, we can estimate the particle size of PR772 virus to be about 70 nm.

Discussion

With intense competition from single-particle cryo-EM [29], the current development of SPI is limited by two factors — particle hit rate and sample heterogeneity. As discussed in Section 3.1.2, we should expect to collect at least 10^5 diffraction patterns of single particles to achieve a half-period resolution of 14 Å. If the injected particles manifest structural heterogeneity, for example, the wide particle size distribution due to the aggregation of non-volatile contaminants around the injected particles [14], more diffraction patterns would be required to reconstruct the structures of different structural classes.

The hard X-ray beamline, CXI, of the LCLS allows data to be collected at angstrom resolution. However, the small beam focus size that counteracts the smaller scattering cross sections at shorter X-ray wavelength results in a low particle hit rate. On the other hand, the soft X-ray beamline, AMO, of the LCLS produces a larger beam focus size and hence allows higher particle hit rate, but the resolution is limited to several nanometers due to the physical limitations of beamline design. Although the European XFEL and the upcoming upgrade of LCLS II will increase the X-ray pulse repetition rate by three orders of magnitude to greatly increase the number of particle hits, significant advances in injector technology are necessary to make SPI a feasible high-resolution technique in structural biology.

CHAPTER 4

TABLE-TOP SPARSE CRYSTALLOGRAPHY

With SPI at sub-nanometer resolution still beyond our reach, the most successful technique developed at XFELs so far is arguably serial femtosecond crystallography (SFX) [10, 11]. In SFX experiments, data frames are collected from protein nanocrystals[1] sequentially delivered in random orientations into the X-ray beam, which avoids the need to grow large crystals in conventional crystallography experiments. The femtoseconds long pulses allow the photon scattering process to outrun the radiation damage of the crystals, and the high fluence of the pulses enables sufficient photons to be scattered to a fast-framing detector [59] to determine the crystal orientations by indexing individual data frames. The protein structure is solved by merging the crystal diffraction patterns in reciprocal space and phasing the resulting Fourier magnitudes.

Although developments in detector technology, sample delivery and data analysis have made SFX a viable technique, its wide use is limited by the scarcity of XFEL beamtime. Despite the construction of XFELs worldwide, available beamtime at XFELs will still be scarce compared to that provided by the existing storage ring synchrotron sources in the near future. This has inspired development of serial microcrystallography experiments at current storage ring sources [9, 27, 29, 32, 48, 55, 62, 69], where protein structures are solved by merging diffraction patterns of many unoriented, individual microcrystals. Since the pulse width of storage ring sources is of the order of picoseconds, radiation damage cannot be outrun in the same way as at XFELs. At storage rings the exposure time per crystal is limited by radiation damage. If the crystal is too small, too few X-rays to determine the crystal orientation will be diffracted prior to irreversible radiation damage. Therefore, serial crystallography at storage ring sources has thus far

[1]The term 'nanocrystal' has been loosely used to refer to crystals of sizes ranging from a few hundred nanometers to several micrometers.

relied on relatively large crystals. Frames with insufficient resolvable Bragg peaks for indexing, which we call 'sparse frames', are simply discarded. Proteins not bound up in large crystals are wasted for the purpose of structure determination.

Using the EMC algorithm, we have developed an alternative analysis method that makes use of the sparse frames without exceeding a dose that would damage the crystal. Unlike indexing algorithms that determine a definite orientation on a per frame basis, the EMC algorithm models the orientation of each frame probabilistically and reconstructs a consistent 3D intensity model using all the data frames simultaneously. The information from a sparse frame still contributes to the reconstruction even though the frame alone cannot be indexed. This approach can reduce the usable crystal size in SMX experiments at storage ring sources and extract information from the sparse frames that would otherwise have been discarded.

In this chapter, we demonstrate the ability of the EMC algorithm to handle sparse frames with two proof-of-concept experiments. In these experiments, large protein crystals were illuminated by a dim lab X-ray source to simulate sparse frames collected from microcrystals at storage ring sources. By increasing the experimental complexity, the EMC algorithm has been developed to take on the analysis of a real SMX dataset collected at a storage ring source, which is the focus of the next chapter. The contents of this chapter have been published in Ref. [41] and [79].

4.1 Single-axis data

This study is part of a methodical program that aims to analyze sparse crystal diffraction data collected in SMX experiments at storage ring sources. In Ref. [58] and [4], it is shown that the probabilistic modeling of the EMC algorithm continues to hold even

with just a few photons per frame in 2D and 3D shadowgraphy. In Ref. [5], the EMC algorithm was used to reconstruct the 3D intensity map from sparse frames collected from a small-molecule crystal rotated about a single axis. Here we show a successful 3D intensity reconstruction from sparse frames without any resolvable Bragg peaks, which were collected from a protein crystal rotated about a single axis. It is further demonstrated that the protein structure can be solved from the reconstructed Bragg intensities.

4.1.1 Data collection

In our first proof-of-concept experiment [79], a single hen egg white lysozyme (HEWL) crystal of approximately 400 μm in size was mounted on a goniometer and set continuously rotating on a rotation stage at 0.05° per second, with the rotation axis set to be perpendicular to the incident beam. The crystal was illuminated by a Cu K_α X-ray beam (1.54 Å in wavelength) generated by a rotating anode X-ray generator. A cryostream was used to maintain the crystal at 100 K to help protect it from radiation damage. The X-ray beam had a flux density of 40 photons $\cdot$ μm^{-2} $\cdot$ s^{-1} and a divergence of 1 mrad. The data frames were recorded by a fast-framing Mixed-Mode Pixel Array Detector (MM-PAD) [71] at a distance 33 mm from the crystal. The center of the beam was placed in one corner of the active area of the MM-PAD to record the highest possible resolution[2], which was approximately 1.3 Å. A PIN-diode beamstop was used to keep the direct beam from striking the detector. The schematic of the experiment is shown in Figure 4.1.

[2]With wavelength λ and scattering angle θ, crystallographers define the resolution by

$$(\Delta x)_{\text{full}} = 2\pi/q = \frac{\lambda}{2}\sin^{-1}\frac{\theta}{2}, \tag{4.1}$$

where $(\Delta x)_{\text{full}}$ is also called the full-period resolution and is twice the half-period resolution.

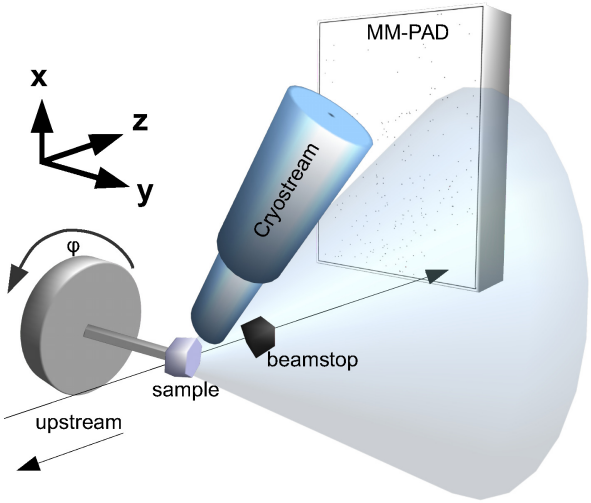

Figure 4.1: Schematic of the single-axis sparse crystallography experiment (not drawn to scale). The X-ray beam is incident from the left side of the image along the z-axis, with the crystal rotated about the y-axis. A cryostream cools and maintains the crystal at 100 K. The diffracted photons are recorded by the MM-PAD, and the main beam is blocked by a beamstop.

We ensured data sparsity by reducing the exposure time per frame to a sufficiently short duration. An exposure time of 10 ms was used, which corresponds to a 0.0005° oscillation angle per frame. A total of 8.8 million frames were collected (12 full revolutions of the crystal), with about 200 photons per frame on average (Figure 4.2).

4.1.2 Data analysis

EMC reconstruction

We sampled the rotations uniformly about the single rotation axis, whose orientation in the crystal reference frame was obtained by merging the data frames collected in the first crystal revolution into bins of size 1° and indexing with the XDS package [37]. Using the lattice parameters estimated by indexing, the initial 3D intensity model was seeded by placing small 3D Gaussians of random height at each predicted Bragg position. No

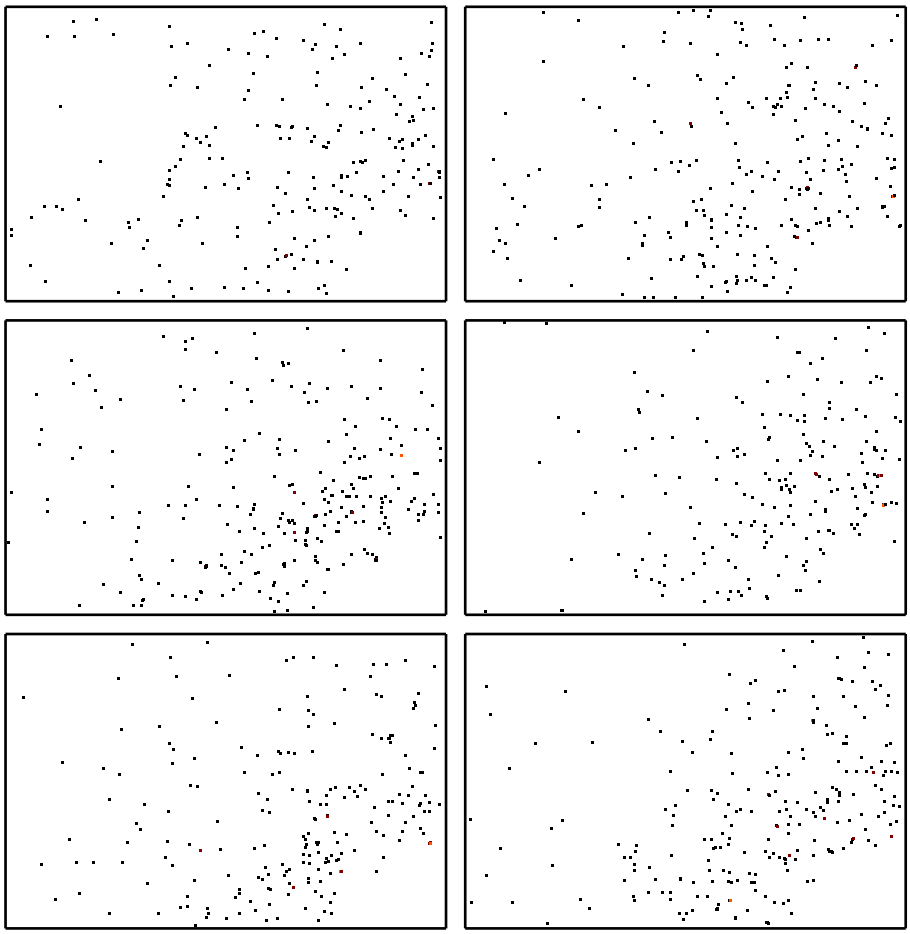

Figure 4.2: Random selection of six data frames (262×393 pixels). The direct beam is incident normally in the lower right region of the detector, which is blocked by the beamstop. The resolution at the upper left corner is 1.3 Å. Each frame consists of only 200 photons on average and the maximum photon count in these frames is three per pixel. The size of the pixels is smaller than the rendered photons in this image, which are enlarged for visual clarity.

symmetry, such as Friedel pairs or systematic absences, was imposed. We note that this initialization step was the only time that information about the relative angles between data frames was used.

The 3D intensity model, $W(\mathbf{q})$, was reconstructed using the standard EMC algorithm described in Section 2.3.1. Since the lab X-ray source is unpolarized, p_i, the product of

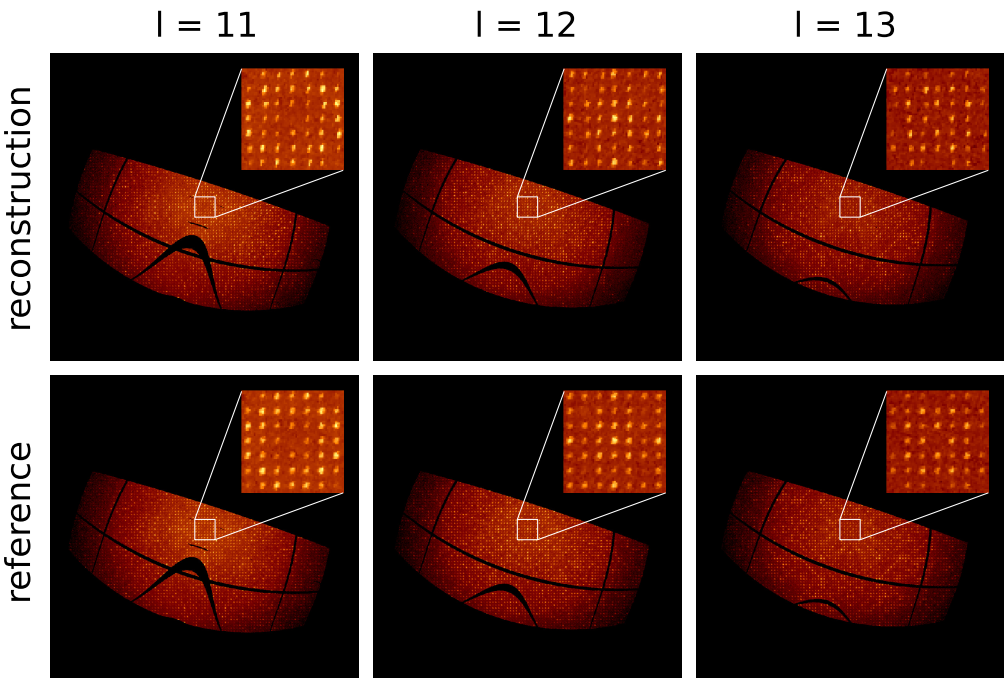

Figure 4.3: Slices of the reconstructed and reference intensity models in the hk plane at constant values of l. Even without imposing symmetry, the reconstructed intensity obeys the reflection condition $00l : l = 4n$ required by the $P4_32_12$ space group symmetry of the HEWL crystal (see insets). The mapping to reciprocal space transforms the detector gaps [71] into curves.

the polarization factor and solid angle for pixel i, is a constant at fixed spatial frequency magnitude, q. The factor p_i can hence be absorbed into the definition of $W(\mathbf{q})$, and would be divided out from the reconstructed Bragg intensities before solving the structure. The photon count, K_{ik}, measured by pixel i in data frame k is the Poisson sample of the mean photon number, W_{ij}, measured by pixel i given crystal orientation j. No symmetry was imposed in the reconstruction. The EMC iterations continued until the 3D intensity model, $W(\mathbf{q})$, converged. On convergence, we rescaled the values of $W(\mathbf{q})$ so that the sum of its voxel values equalled the total number of photons recorded in the dataset. By Poisson statistics, the variance of each voxel is then given by the voxel value.

The reconstructed intensity model was compared with the actual intensity model. The actual (i.e. 'reference') model was constructed using the known orientation of each

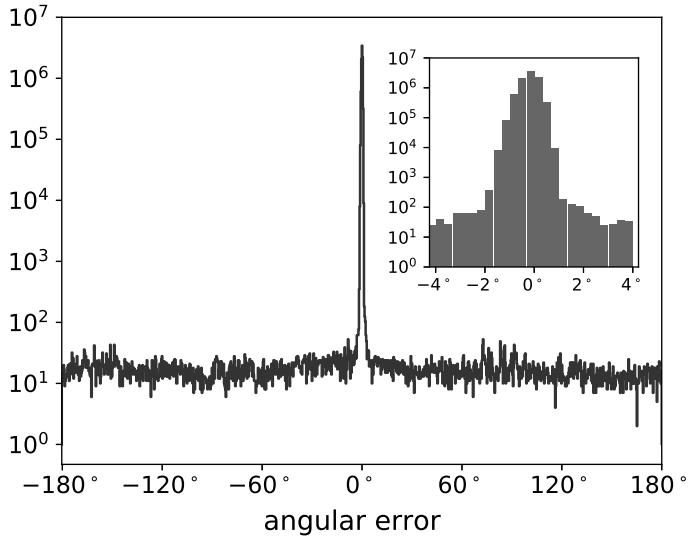

Figure 4.4: Histogram of the difference between the reconstructed most probable orientation and the actual orientation for each frame, expressed in degrees about the rotation axis. The EMC algorithm correctly assigned 99.7% of the frames within 1°, as shown in the inset.

frame, even though this information was not used in the EMC reconstruction. Several slices of the reconstructed and reference intensity models perpendicular to the l-axis of the reciprocal lattice are shown in Figure 4.3. The reconstructed intensity obeys the reflection conditions (structure factors not systematically zero) $00l : l = 4n$ and $h00 : h = 2n$ required by the $P4_32_12$ space group symmetry of the HEWL crystal [30]. Since no symmetry was imposed in either the seeding or reconstruction process, this suggests a successful reconstruction.

A more direct validation of our reconstruction is the difference between the most probable orientation of each frame assigned by the EMC algorithm and its actual orientation, which is shown in Figure 4.4 as a histogram of relative angles about the rotation axis. We found that 99.7% of the frames were assigned to the correct orientation within 1°. The outliers are possibly caused by an abnormally low SNR in some data frames, for example, frames recorded at crystal orientations with few Bragg spots intercepted by the Ewald sphere, or frames that suffered extra background scatter from the sample holder.

This shows the importance of background reduction in future experiments, specifically in the case of small or weakly diffracting crystals.

Integration of Bragg peaks

Since the EMC algorithm placed no special focus on the Bragg peaks, everything present in the data frames — background, diffraction spots, diffuse scatter, etc. — were reconstructed. In order to extract the information of the Bragg intensities, we used a 3D version of the peak-segmentation algorithm described in Ref. [81]. The segmentation is a classification of voxels into signal or background based on a z-score

$$z(w) = \frac{w - \mu}{\sigma}, \tag{4.2}$$

where w is the value of the voxel in consideration, and μ and σ represent the mean and standard deviation of the values of background voxels in a surrounding $n \times n \times n$ cube. Voxels with z-score above a particular threshold, γ, are classified as signal; otherwise they are considered as background. In the first iteration, all voxels were used to calculate μ and σ. After that, only voxels classified as background in the previous iteration were included. For good-quality segmentation of the Bragg peaks, we gradually increased γ from 1.0 to 3.0 in successive iterations.

Using the segmented Bragg peaks, we refined the estimates of lattice parameters. For a candidate set of lattice parameters, we computed the total intensity of segmented peaks lying within ellipsoids centered on the corresponding Bragg positions. The ellipsoid volume was a small fraction of the reciprocal unit cell, with principal axes consistent with the tetragonal cell. The lattice parameters giving the greatest total intensity were taken as the refined values. At each predicted Bragg position, an ellipsoid window of a larger volume was used for peak integration. If a voxel is within such a window, it is

assigned to the corresponding peak; otherwise, it is classified as background. The mean of the background voxels surrounding each Bragg position was subtracted from each signal voxel before the signal voxels were summed to give the Bragg intensities, whose variances were calculated using error propagation. Partial peaks, such as those adjacent to boundary, detector gaps or the beamstop region, were rejected. Finally, the corrections due to polarization factors and solid angles were applied to the Bragg intensities and variances, from which we calculated the structure factor magnitudes and their variances.

Structure solution

In order to retrieve the phase information, we input the reconstructed structure factor magnitudes and variances to MOLREP [73] from the CCP4 suite [80] to produce a molecular-replacement solution using a template HEWL structure (PDB entry: 193L [75]). The solution was then refined through 20 iterations in REFMAC [50] with both rigid-body and restrained refinements, and was rebuilt in Coot [20] with cyclical refinement.

The structure solved from the EMC-reconstructed intensities agrees well with the template structure, PDB entry: 193L (Figure 4.5). The root-mean-square deviation (r.m.s.d.)

Reconstruction	
Space group	$P4_32_12$
Lattice parameters (Å)	a = b = 77.5, c = 36.2
Resolution (Å)	54.8 – 1.50
Completeness (%)	92.0
Reflections	16,056
Refinement	
Atoms	1,963
$R_{\text{work}}/R_{\text{free}}$ (%)	28.2/32.0
R.m.s.d. for bonds (Å)	0.0192
R.m.s.d. for angles (°)	0.120

Table 4.1: Refinement statistics of the structure solved from the single-axis dataset.

of the C_α atoms between the two structures equals 0.27 Å, which could be attributed to different solvent content during crystallization and water placement during refinement. With the refinement statistics shown in Table 4.1, our structure reconstructed from sparse data compares favorably with structures obtained by more conventional means.

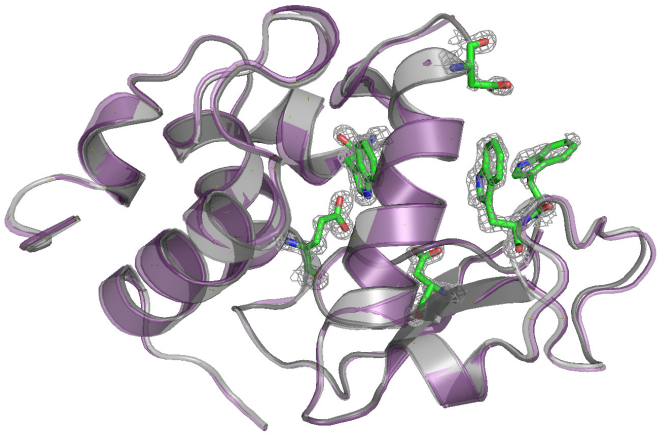

Figure 4.5: Reconstructed protein structure (grey) superimposed on the model (PDB entry: 193L, purple) used in molecular replacement. High resolution features (active sites) are rendered as green sticks (model structure) and grey mesh (reconstruction).

4.1.3 Discussion

In this study, we have shown experimentally that a series of unoriented, sparse diffraction patterns collected from a protein crystal rotated about a single rotation axis can be assembled into a 3D intensity map using the EMC algorithm. The validity of the reconstruction is supported by the recovery of symmetries which were absent in the seeding process and the small angular error in the reconstructed most probable orientation for each frame. Moreover, we have demonstrated that the protein structure can be solved by phasing the reconstructed structure factor magnitudes through molecular replacement. This result suggests that the indexability on a per frame basis does not necessarily limit structure determination in SMX.

Several features are still needed to extend the result of this study to real SMX experiments. The first one is to sample the entire rotation space, which makes the reconstruction much more computationally intensive, and requires significant developments of the EMC algorithm. Another challenge is the background reduction. By counting the number of photons that were not beneath the Bragg peaks in our reconstructed 3D intensity model, we found that 80% of the photons in the dataset came from background scatter, which resulted from air, the solvent surrounding the crystal, and the sample holder. When the diffraction patterns are collected from multiple microcrystals, reducing background photons scattered from the sample delivery medium will be a more critical issue for the success of reconstruction.

4.2 Two-axis data

In order to sample a larger subset of the rotation space, we collected sparse diffraction patterns from a large HEWL crystal rotated continuously about two orthogonal axes in our second proof-of-concept experiment [41]. The local update scheme of the EMC algorithm was developed to speed up the high-resolution reconstruction by two to three orders of magnitude. We have again shown that the crystal intensity can still be reconstructed even without knowledge of the crystal orientation in any sparse frame.

4.2.1 Data collection

The X-ray diffraction patterns were collected from a single HEWL protein crystal centered at the intersection of two orthogonal rotation axes at room temperture. The crystal was illuminated by a Cu K_α X-ray beam (1.54 Å in wavelength) generated by a rotating anode X-ray generator, with a divergence of 1 mrad and a flux density of

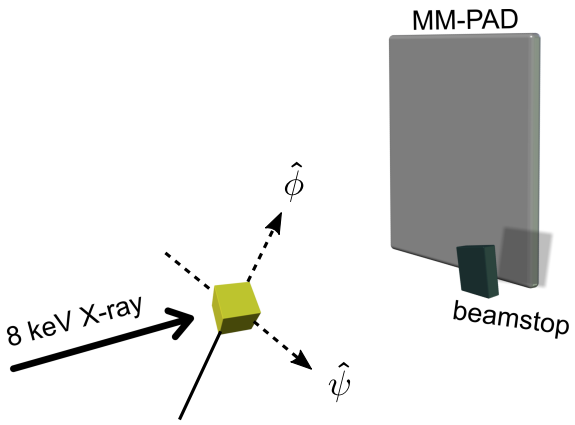

Figure 4.6: A simplified schematic of the experimental setup with two orthogonal rotation axes. The beam incidence is perpendicular to the ψ axis and the MM-PAD, and the main beam is blocked by the beamstop. The crystal is rotated in increments of 0.1° about the ψ axis, with the data frames recorded by the MM-PAD when ϕ traverses 360° continuously at each value of ψ. The figure is not drawn to scale.

40 photons $\cdot\, \mu\text{m}^{-2} \cdot \text{s}^{-1}$. The beam incidence was normal to the MM-PAD and one of the rotation axes. The sample-to-detector distance was 60 mm. The center of the beam was placed in one corner of the active area of the MM-PAD, giving a resolution of 2.0 Å in the opposite corner. A pin-diode beamstop was used to prevent the direct beam from striking the MM-PAD during data collection. The schematic of the experiment is shown in Figure 4.6.

The crystal was rotated about the ψ axis from 0° to 17.9° and then from −18.0° to −0.1° in increments of 0.1°. At each value of ψ, the crystal was rotated by 360° about the ϕ axis continuously at a constant angular velocity of 0.5° per second. The MM-PAD collected images at a framing rate of 4 ms per frame in each revolution of ϕ, which gave an oscillation angle of 0.002° per frame. Owing to radiation damage and possible dehydration of the crystal, we only kept the data frames recorded at ψ ranging from 0° to 15.9° to pass on to processing. This subset of data was chosen by monitoring the decay of high-resolution peaks in the merged diffraction patterns of bin size 1° in ϕ at each value of ψ. We also discarded frames that did not record any photons, which was

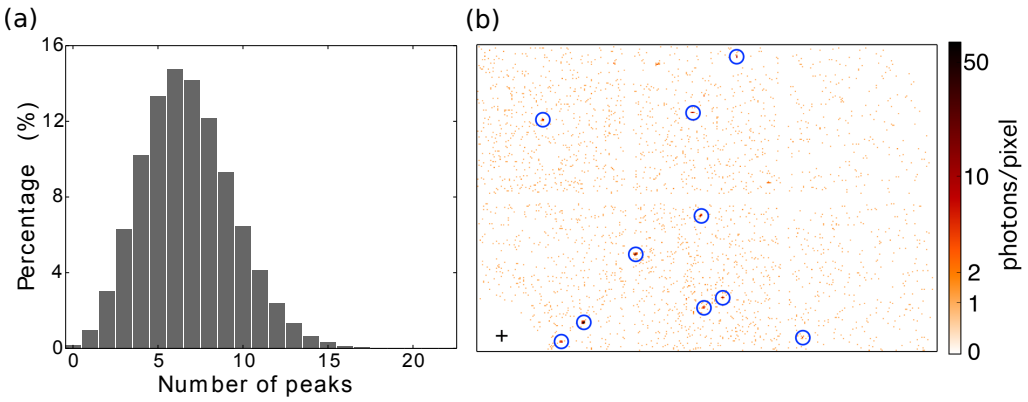

Figure 4.7: (a) Histogram of the number of peaks per collapsed frame, which is the sum of 100 successive frames in the raw data. A cluster with more than two contiguous pixels and at least two photons per pixel on average is identified as a peak. (b) A random selection of the collapsed frames, with the identified peaks marked by blue circles. The cross denotes the beam center, and the resolution at the upper right corner is about 2 Å.

possibly caused by glitches of the rotating anode.

To simulate the signal level of an SMX experiment, we collapsed every 100 successive frames that did not contain any discarded frames, since the data was recorded when the crystal was rotated continuously in ϕ at a fixed value of ψ. The collapse of every 100 successive frames gave us 2.7×10^5 frames with an average of 3000 photons per collapsed frame. By defining a possible Bragg peak as a cluster with more than two contiguous pixels and at least two photons per pixel on average, we obtained the statistics of the number of peaks in each collapsed frame (Figure 4.7). Even with this generous criterion for peak finding, most of the collapsed frames do not have enough peaks to be indexed by conventional means (at least 20 to 30 resolvable peaks per frame).

Following the calculation in Ref. [33], we estimated the energy absorbed by our HEWL crystal over the exposure of one collapsed frame, assuming that protein crystals have the same mass energy absorption cross section as water. Our calculation showed that an 8 μm^3 protein crystal would have endured a 0.2 MGy radiation dose if it had scattered the same number of photons as our large HEWL crystal during this period.

This dose is within the lifetime of protein crystals at room temperature if the radiation is delivered quickly [56], so the signal level in our study should be comparable to that in a real SMX experiment.

It was discovered after data had been collected and the apparatus disassembled that the crystal was of poor quality. We found that even using the known orientations, the resulting structure factor magnitudes cannot be phased to produce a high-resolution structure. The goal of this study, however, was to extend the EMC approach to a greater rotation subset spanned by the two-axis rotations. Because the quality of the reconstructed intensities can be assessed by comparing with the actual intensities, the goal of the experiment could be met even though the crystal was of poor quality for solving the structure.

4.2.2 EMC reconstruction

We first determined the orientation of the crystal reference frame relative to the lab frame by mapping the collapsed data frames to reciprocal space with their known relative orientations to form a 3D intensity map. The reciprocal lattice of the crystal is embedded in the intensity map and differs from the lab frame by a global rotation R_g, which was obtained by segmenting out the Bragg peaks [79] and then indexing the peaks [70]. The intensity map was then rotated by R_g to align with the lab frame, and this aligned intensity map is what we call the reference intensity map.

We generated the discrete rotation samples using the 600-cell subdivision method [46], where the angular resolution $\delta\theta = 0.944/n$ is specified by the order $n = 1, 2, 3, \ldots$. Here we confined the rotation samples to those in the rotation subset explored by the rotated crystal. The range of the subset in the lab frame was given

by applying the global rotation, R_g, to the relative orientations between the collapsed frames. The 3D crystal intensity map was reconstructed from the collapsed frames using rotation samples in this subset, with the orientation of each frame unknown to the EMC algorithm. This choice of rotation samples makes the solution to the two-axis problem directly applicable to the randomly oriented frames in real SMX experiments, where the rotation subset is replaced by the whole 3D rotation space.

The initial 3D intensity map was seeded with small 3D Gaussians of random height at each predicted Bragg position, with the lattice parameters given by the indexing process mentioned above. In real SMX experiments, where the true orientation of each frame is unavailable, the lattice parameters should be estimated by other means, for example, indexing the 1D pseudo-powder pattern, which is the histogram of the identified peaks in all data frames over spatial frequency magnitudes. No symmetry was imposed in either the seeding or reconstruction process.

Given the 2.7×10^5 collapsed frames, we started an EMC reconstruction using the standard update scheme described in Section 2.3.1. As in the single-axis case, the polarization factors and solid angles were absorbed into the definition of $W(\mathbf{q})$. The measured photon count, K_{ik}, was modeled by the Poisson sample of the mean photon number, W_{ij}. Rotation samples of order $n = 40$ and data up to 3 Å resolution were used in this stage to quickly obtain a converged low-resolution reconstruction. After the convergence was reached, a high-resolution 3D intensity map was constructed using data up to 2 Å resolution and the probability distribution of orientations. This intensity map was then used as the initial model for the local update scheme (Section 2.3.3) using rotation samples of orders $(n_c, n_f) = (40, 60)$ for refinement. Different pairs of orders, (n_c, n_f), with increasing angular resolutions were sequentially used in the local update scheme to extend the peak convergence to high resolution. The converged intensity map

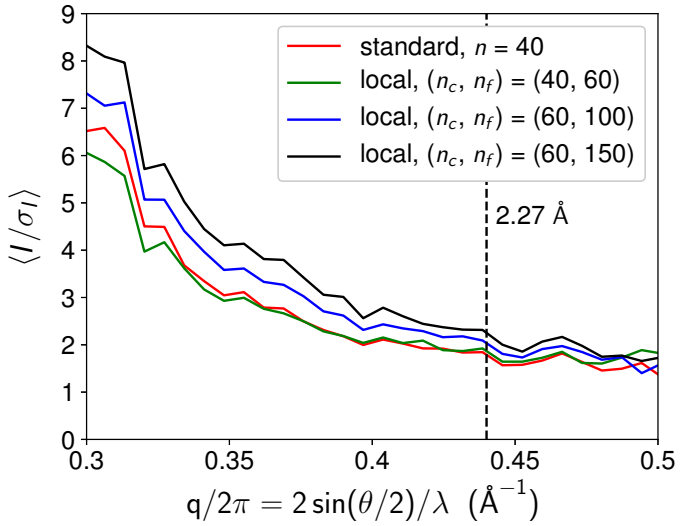

Figure 4.8: The average SNR of the integrated Bragg intensities from the converged intensity maps at different stages of the reconstruction. The increase of $\langle I/\sigma_I \rangle$ at high resolution indicates the reconstruction of high-resolution peaks. The 2.27 Å resolution determined by CC^* (see below) is marked by the black dashed line.

was subsequently rescaled so that the sum of voxel values equalled the total number of photons collected in the dataset. Following the procedure of peak integration in the single-axis case, we obtained the Bragg intensities and their variances. Partial peaks, such as those adjacent to boundary, detector gaps or the beamstop region, were rejected.

Figure 4.8 shows the average SNR of the integrated Bragg intensities, $\langle I/\sigma_I \rangle$, from the converged intensity maps at different stages of the reconstruction. In the transition from the standard update scheme of $n = 40$ to the local update scheme of $(n_c, n_f) = (40, 60)$, the values of $\langle I/\sigma_I \rangle$ dropped at low resolution but remained at similar levels at high resolution. The lack of improvement at high resolution indicates that the current angular resolution of the local update scheme still cannot resolve high-resolution peaks. On the other hand, the inclusion of data beyond 3 Å slightly disrupted the original probability distribution, which in turn reduced $\langle I/\sigma_I \rangle$ at low resolution. The improvement of $\langle I/\sigma_I \rangle$ when increasing the angular resolutions shows the reconstruction of high-resolution peaks and justifies the local update scheme.

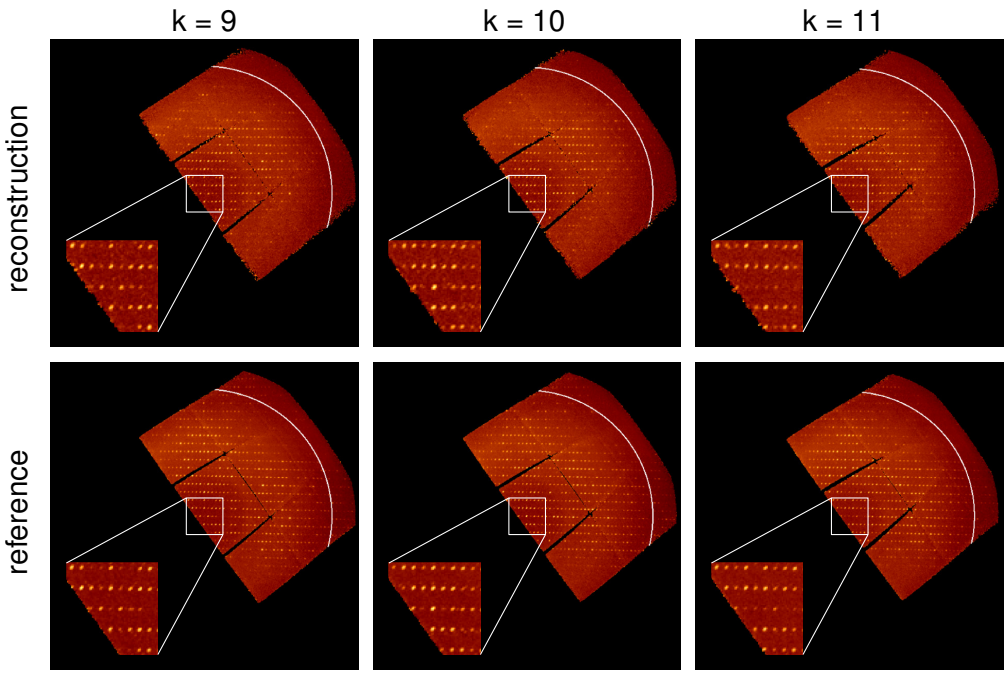

Figure 4.9: Slices of the reconstructed and reference intensity maps in the hl plane at constant values of k of the reciprocal lattice. Even without imposing any symmetry in either the seeding or reconstruction process, the converged intensity map still follows the reflection conditions required by the space-group symmetry $P4_32_12$ of the HEWL crystal (see insets). The 2.27 Å resolution determined by CC^* is marked by the arcs in white. The mapping into reciprocal space transforms the detector gaps [71] into curves.

With the converged intensity map from the local update scheme of $(n_c, n_f) = (60, 150)$ as our final intensity reconstruction, Figure 4.9 compares the slices of the reconstructed and reference intensity maps perpendicular to the k-axis of the reciprocal lattice. As in the single-axis case, the recovery of symmetries in the Bragg intensities demonstrates the success of the EMC reconstruction. We note that the discrepancy between the two intensity maps in high-resolution peaks is consistent with the low SNR at high resolutions (see Figure 4.8). Since the photons contributing to the high-resolution shells were mostly collected by the upper left corner of the MM-PAD (Figure 4.6), the resulting lower SNR made the orientation reconstruction more challenging in this region.

A further comparison is shown in the scatter plot of the integrated Bragg intensities from the reconstructed and reference intensity maps (Figure 4.10), which excludes the

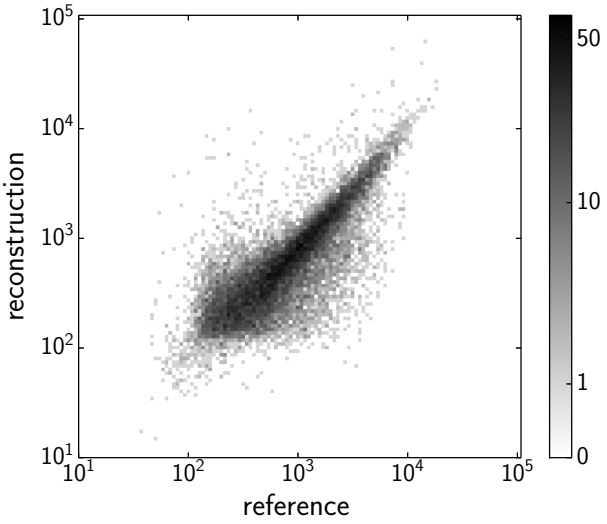

Figure 4.10: Scatter plot comparing the integrated Bragg intensities from the reconstructed and reference intensity maps. Integrated intensities with SNR $I/\sigma_I < 2$ are excluded from the plot. The linear correlation shows the agreement between the two intensity maps.

reflections with SNR $I/\sigma_I < 2$. The linear correlation of the integrated intensities shows the consistency of the two intensity maps. By summing the total photon counts of both the integrated and the partial peaks, we also estimated the fraction of photons coming from the background and diffuse scatter as about 90%, which was mainly scattered by air, the solvent surrounding the crystal, and the sample holder.

To estimate the resolution of our reconstruction, we calculated the correlation coefficient of the observed reflections with the underlying true signal, CC^*. We first randomly separated the symmetry-related peaks into two halves, and calculated the unique reflections by averaging the symmetry-related peaks in each half. The correlation coefficient between the unique reflections of the two halves, $CC_{1/2}$, was then computed in different resolution shells. Under the assumption that the errors of the two halves are independent, identically distributed and free from the errors of the true signal, the value of CC^* is given by [38]

$$CC^* = \left(\frac{2CC_{1/2}}{1 + CC_{1/2}}\right)^{1/2}. \tag{4.3}$$

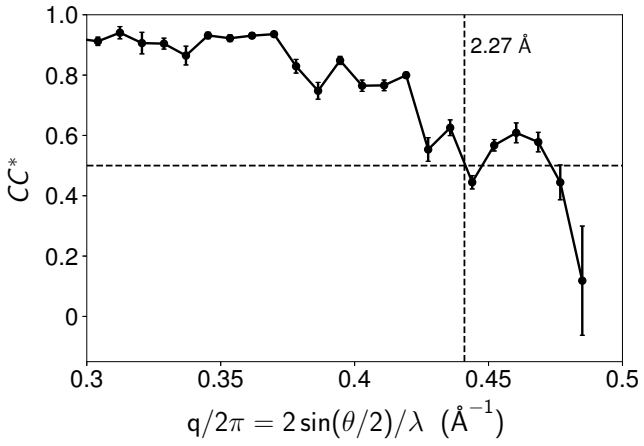

Figure 4.11: Plot of CC^* as a function of spatial frequency magnitude. The resolution of the reflections is determined as 2.27 Å by a threshold $CC^* = 0.5$. The error bars are estimated by repeating the random separation of symmetry-related peaks 1,000 times, while the ups and downs in CC^* result from the binning in resolution shells.

The plot of CC^* as a function of spatial frequency magnitude is shown in Figure 4.11, with the error bars estimated by repeating the random separation of symmetry-related peaks 1,000 times. The large error bar in the highest-resolution shell shows the low correlation between the integrated intensities of the two halves, which is consistent with the low SNR at high resolution. The resolution of the reconstructed Bragg intensities was determined as 2.27 Å by the threshold $CC^* = 0.5$. We note that the value of CC^* is dominated by the stronger peaks in each resolution shell. Therefore, CC^* can still have moderate values even if some weak peaks are not resolvable, as indicated by the discrepancy between the two intensity maps in high-resolution peaks in Figure 4.9.

4.2.3 Discussion

Here we have demonstrated the 3D intensity reconstruction using the EMC algorithm from the sparse diffraction patterns collected from a large HEWL crystal. The crystal was rotated about two orthogonal axes to sample a greater subset of the rotation space.

To address the increased computational load, we developed the local update scheme of the EMC algorithm to speed up the high-resolution reconstruction. These developments have brought us one step closer to the goal of applying the EMC approach to reduce the usable crystal sizes in current SMX experiments.

In this study we used a large single crystal rotated in various orientations to emulate the data expected from multiple small crystals. The obvious next step towards practical application of the method is to try the EMC algorithm on data collected from multiple small crystals. It will be necessary to experimentally determine the severity of difficulties arising from sources including varying crystal diffraction quality and occasional multiple crystals in the beam. In addition, the EMC algorithm also needs to calculate the frame-to-frame signal strength variation arising from crystal size variation. These issues together with improved estimates of background scatter are addressed in the next chapter to determine a 3D protein structure from an SMX dataset collected at a storage ring source.

CHAPTER 5

SERIAL MICROCRYSTALLOGRAPHY AT A STORAGE RING SOURCE

Adapted from SFX experiments at XFELs, most room-temperature SMX experiments carried out at storage ring sources also adopt the same data analysis pipelines as SFX. When it comes to crystal intensity reconstruction, the most widely used approach is the combination of the packages, *Cheetah* [7] and *CrystFEL* [77]. *Cheetah* consists of a set of high-throughput data reduction programs for serial diffraction patterns. It identifies possible crystal diffraction patterns by a threshold on the number of resolvable peaks per frame. These patterns are passed to *CrystFEL* to determine the crystal orientations by indexing methods. The Bragg intensities are subsequently obtained by the Monte Carlo integration method [39], which calculates the average of the indexed peak values after background subtraction and corrections for polarization and solid angle.

Here we describe an alternative approach that uses the EMC algorithm to reconstruct the Bragg intensities from SMX data collected at storage ring sources. This approach is demonstrated on an SMX dataset graciously provided by the authors of Ref. [48]. In particular, we threw away the strong crystal diffraction patterns and focused on the data frames that cannot be indexed by conventional means. Despite the daunting background scatter from the sample delivery medium, we still managed to solve the protein structure at 2.1 Å resolution. In contrast to the Monte-Carlo integration approach, our method uses the reconstructed crystal volumes, for all the data frames, when building the 3D intensity model. By lifting the requirement of indexability for each data frame, protein structures can be determined with small or weakly-diffracting crystals at storage rings. The contents of this chapter will appear in Ref. [42].

5.1 Data reduction

The SMX dataset we used was collected by Martin-Garcia and coworkers at the GM/CA 23-ID-D beamline at the Advanced Photon Source [48]. The raw data consists of 304,643 frames[1] measured from HEWL microcrystals of size ranging from 5 to 10 μm at room temperature. The crystals were sequentially delivered to the X-ray beam in random orientations by a lipidic cubic phase (LCP) gel injector with a glass nozzle of 50 μm inner diameter [76]. In order to demonstrate the ability of our method to handle weak crystal diffraction data, we excluded the data frames with more than 20 resolvable Bragg peaks (see below), which is the empirical lower bound for normal indexing methods to succeed. In other words, we only consider the weak crystal diffraction patterns that were rejected for the structure determination in Ref. [48], which amounts to 120,574 sparse frames.

The data reduction started by identifying the frames containing crystal diffraction signals because the crystals were randomly distributed in the LCP gel. This process, also known as 'hit finding', first locates possible Bragg peaks in the diffuse background scatter. Our method is based on outlier detection. In the absence of crystal diffraction, the probability that a pixel measures a photon count, K, follows the Poisson distribution, $P_b(K) = e^{-b}b^K/K!$, where b is an estimate (described below) of the photon number at that pixel due to the diffuse background scatter. Given b, we can identify an outlier pixel by its photon count being too large to be consistent with Poisson statistics. This consistency is defined via a photon count threshold, $\tilde{K}$, defined by the cumulative probability:

$$\min_{\tilde{K}} \sum_{K=0}^{\tilde{K}} P_b(K) > 1 - \epsilon, \tag{5.1}$$

[1]This dataset is a representative subset of the data reported in Ref. [48] (364,724 frames in total), without any pre-selection.

where ϵ is a small number that lets us set a false-positive rate (see below). If the photon count measured in the pixel exceeds the threshold, $\tilde{K}$, we assume that crystal diffraction contributed to the signal.

Since we had no prior knowledge of the background photon numbers, b, we estimated them by the following self-consistent iterative scheme. Observing that the background scatter is generally azimuthally symmetric about the incident X-ray beam, we assumed that b only depends on the frame index, k, and the spatial frequency magnitude, q. The initial values of b_{qk} were obtained by averaging all photon counts in annular regions, after the pixel-wise correction of the polarization factor and solid angle. Because the number of pixels in these annular regions ranged from 10^3 to 10^4, the value of ϵ in Equation (5.1) was set to 10^{-5} to reduce false positives arising from statistical fluctuations. In each iteration we used the current estimates of b_{qk} to calculate the pixel-wise background estimates, b_{ik}, by the relation

$$b_{ik} = p_i b_{qk}, \tag{5.2}$$

where p_i is the product of the (positive) polarization factor and solid angle of pixel i. From the values of b_{ik}, we identified the outlier pixels and excluded them from the annular average for b_{qk} in the next round. This procedure was repeated until the values of b_{qk} converged, giving us a good estimate of the background scatter and a list of outlier pixels for each data frame.

The photon count thresholds, $\tilde{K}$, defined by Equation (5.1) with $\epsilon = 10^{-5}$, are plotted in Figure 5.1(a) over a range of background estimates, b. Also shown is the SNR, which is defined as the ratio of $\tilde{K}$ to b. We can see that the SNR takes on a wide range of values over b, especially when the values of b are close to zero. Since the background estimates in the data frames used here range from a fraction to 20 photons, the threshold values defined by the cumulative Poisson probability detects outliers in a more consistent

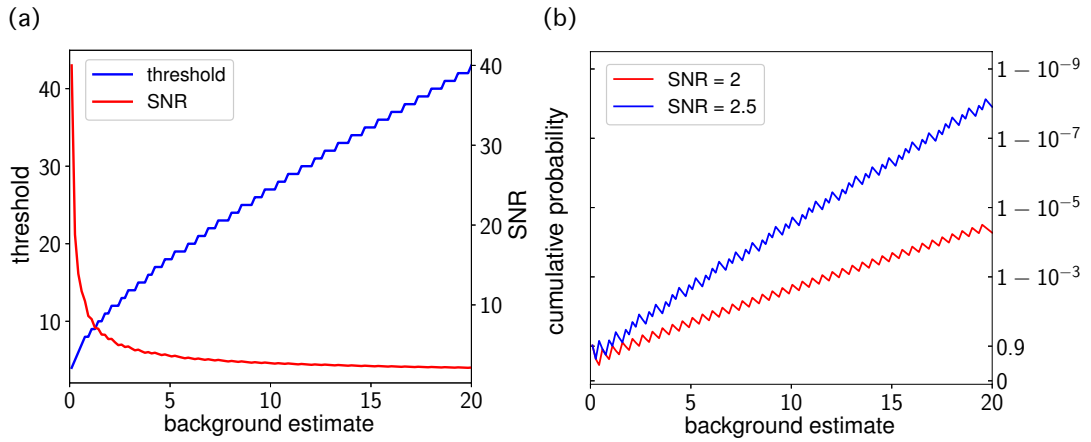

Figure 5.1: (a) The photon count thresholds determined by Equation (5.1) with $\epsilon = 10^{-5}$. The SNR is defined as the ratio of the thresholds to the background estimates. (b) The cumulative probabilities, $P_b(K \leq b \cdot \text{SNR})$, to measure photon count, K, that is no larger than the thresholds, $b \cdot \text{SNR}$, defined by fixed values of SNR over a range of background estimates, b.

way than those determined by a fixed SNR. Figure 5.1(b) further illustrates this point by plotting the cumulative probabilities, $P_b(K \leq b \cdot \text{SNR})$, for different thresholds defined by fixed values of SNR. Under this definition, photon counts greater than the threshold, $b \cdot \text{SNR}$, are identified as outliers, which may result in many false positives at small values of b. In practice, SNR is usually used along with other criteria that characterize a peak in the hit finding process.

We were able to identify Bragg peak candidates as clusters that contain 2 to 10 contiguous outlier pixels, because most clusters have sizes smaller than 5 pixels. Clusters with more than 10 contiguous outlier pixels were considered as originating from something other than Bragg spots and were masked out for the rest of the analysis. As mentioned above, frames with more than 20 candidate peaks were discarded to keep the sparse frames only. Given the Bragg spot locations in the remaining frames, we estimated the lattice parameters by constructing a 1D pseudo-powder pattern as follows. After mapping the candidate peaks to reciprocal space, we recorded the distances

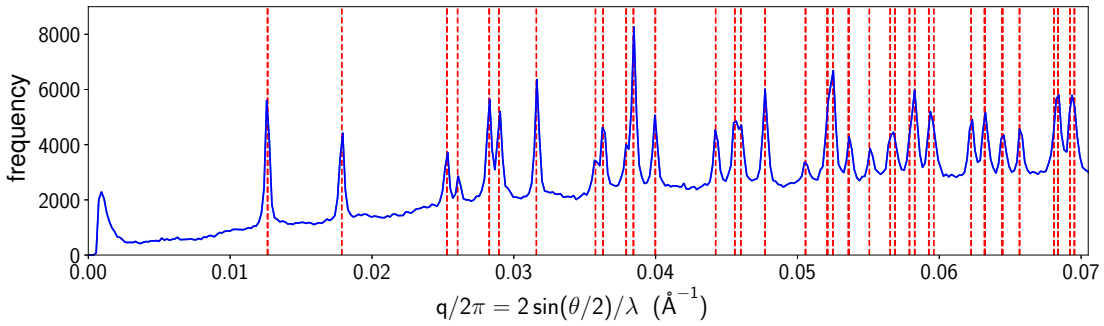

Figure 5.2: The 1D pseudo-powder pattern generated from the frequency of the inter-peak distances in reciprocal space. The red dashed lines indicate the peaks predicted by a primitive tetragonal lattice with lattice parameters a = 79.1 Å and c = 38.4 Å. The peak closest to the origin represents pairs of Bragg peak candidates that are very close to each other. These pairs are actually fragments of Bragg spots of a larger size.

between the centroids of the peaks, for all the data frames. By dividing the spatial frequency magnitudes into bins of the same size, the 1D pseudo-powder pattern was given by the histogram that records the frequencies of the inter-peak distances in each bin. The inter-peak distances are a more reliable source of information about the lattice geometry than the spatial frequency magnitudes of the peaks because the low-resolution peaks are made inaccessible by the beamstop. By assuming a primitive tetragonal lattice to simplify the analysis in this study, the lattice parameters were estimated as a = 79.1 Å and c = 38.4 Å by fitting the peaks in the 1D pseudo-powder pattern (Figure 5.2).

In principle, we should be able to determine the lattice parameters from the 1D pseudo-powder pattern even without the knowledge of the unit cell type. This can be done by an exhaustive search over combinations of lattice parameters from unit cells with high symmetry to those with low symmetry. In the challenging cases of crystals with low symmetry and large unit cell dimensions, we can expect to have a separate measurement of crystal diffraction patterns by moving the detector further away from the interaction point. The 1D pseudo-powder pattern in this case would be the sum of resolvable peak values over spatial frequency magnitudes. Sample consumption should not be a concern

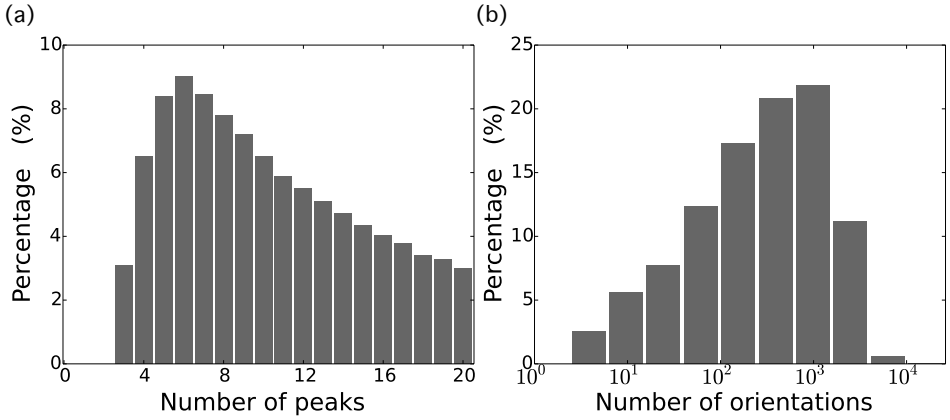

Figure 5.3: (a) The number of candidate peaks in each crystal-hit frame. The data frames with more than 20 peaks were excluded from this study. (b) The number of possible orientations for each crystal-hit frame, which were determined by an exhaustive search in the rotation space using the identified peaks within 4 Å.

here, since the number of peaks needed to populate the 1D pseudo-powder pattern is of similar order to the lattice parameters to be fitted (at most 6). These low-resolution crystal diffraction patterns can also be incorporated to the EMC reconstruction to improve the statistics of Bragg intensities at low resolution.

Finally, we narrowed down the possible crystal orientations per frame by taking advantage of the crystal lattice. The centroids of the candidate peaks within 4 Å resolution in each frame were rotated over all rotation samples. A frame was considered a 'crystal hit' when at least 3 candidate peaks matched the predicted Bragg positions within a predefined distance, r_p, at some orientation, and those frames with no such orientations were simply discarded. The rotations were sampled by the 600-cell subdivision method [46] at order n = 70, which corresponds to an angular resolution of $0.944/n \sim 13.5$ mrad. This procedure reduced the data to 120,574 crystal-hit frames. As shown in Figure 5.3, the possible orientations for each frame are still far from unique.

5.2 EMC reconstruction

We modeled the diffraction pattern of each crystal hit as the Poisson sample from the incoherent sum of the background estimates and the crystal diffraction. For data frame k that records the diffraction of a crystal at orientation j, the mean photon number measured by pixel i is given by

$$\widetilde{W}_{ijk} = b_{ik} + p_i \phi_k W_{ij}, \tag{5.3}$$

where ϕ_k is a scale factor proportional to the crystal volume, the X-ray beam fluence and the travel time of the crystal across the beam, and W_{ij} denotes the value sampled by pixel i from the 3D crystal intensity model, W, at crystal orientation j. The Poisson sample from $\widetilde{W}_{ijk}$ gives the photon count, K_{ik}, with the crystal orientation unmeasured. Our main task is to reconstruct W and ϕ_k given the data, K_{ik}, and the background estimates, b_{ik}, which applies to the experimental condition described by Equations (2.55) to (2.59).

5.2.1 Low-resolution reconstruction

Since the computation time of the EMC algorithm is proportional to the number of pixels and rotation samples, we began with a low-resolution reconstruction. The pixels with resolution higher than 4 Å were masked out in the 120,574 selected frames, and the rotation samples for each frame were limited to the possible crystal orientations recorded in the hit-finding process. All the photon counts within the resolution cutoff were input to the EMC algorithm to reconstruct both the strong and weak intensities. We seeded the 3D intensity model, W, with 3D Gaussians of random height at each Bragg position, and only allowed the voxels within the predefined radius, r_p, about the Bragg positions to be non-zero throughout the reconstruction. The scale factors, ϕ_k, were initialized by the

average value of the identified peaks in each frame. To achieve the highest resolution, we imposed the tetragonal and Friedel symmetries on the values of W after each update to increase the SNR of the Bragg peaks. In general, EMC reconstructions succeed even without imposing symmetries [41, 79].

The values of ϕ_k were held fixed in the first few EMC iterations to rapidly obtain a rough estimate of W. The updates then alternated between W and ϕ until the models converged. Since a data frame may record diffraction signals from multiple crystals in real SMX experiments, these multi-crystal frames have to be rejected to avoid compromising the reconstruction, and this task was completed using the converged probability distribution, P_{jk}. When a data frame has non-negligible probabilities at two orientations, j_1 and j_2, which cannot be related by the crystal point group symmetry, it is likely that the diffraction signals were scattered from two different crystals. We identified 528 such multi-crystal frames and excluded them together with the 2,142 frames with $\phi_k = 0$ from the later analysis. Using the remaining 117,904 single-crystal frames, we updated W for a few more iterations by fixing the values of ϕ_k until the new convergence was reached.

Figure 5.4(a) shows the central slice of the reconstructed 3D intensity model, W, perpendicular to the l-axis of the reciprocal lattice. Each spot represents an integrated peak value in arbitrary units. After dividing the reconstructed values of ϕ_k by the beam fluence and the crystal exposure time, we obtained crystal volume estimates for the single-crystal frames. In order to put these on an absolute scale, we further rescaled their values so that the largest crystal has size of 10 μm, the value reported in Ref. [48]. The resulting crystal volume distribution has 73% of the frames with crystal volume below 100 μm^3 (Figure 5.4(b)). Since our analysis excluded the frames with more than 20 peaks, which generally have larger crystal sizes, this distribution represents the upper limits of the crystal volumes illuminated by the X-ray beam.

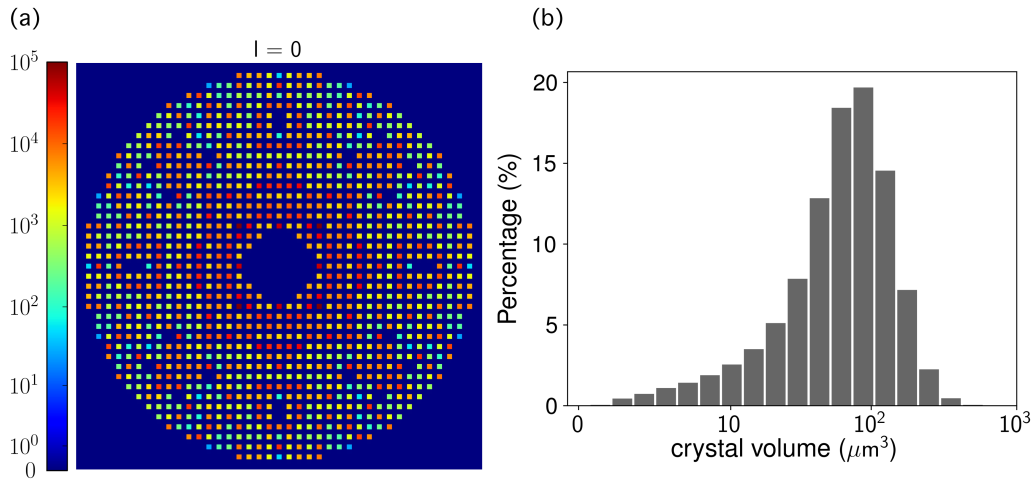

Figure 5.4: (a) The central slice of the low-resolution 3D intensity model, W, perpendicular to the l-axis of the reciprocal lattice. Each spot represents an integrated Bragg peak in arbitrary units, with the negative reflections thresholded to zero for rendering. (b) The reconstructed crystal volume distribution for the single-crystal frames. The values of the crystal volume were rescaled so that the largest crystal size is 10 μm.

5.2.2 High-resolution reconstruction

Based on the low-resolution models, we extended the reconstruction to high resolution using data up to 2 Å resolution. We initialized the 3D intensity model, W, by placing 3D Gaussians of random height at each Bragg position, and replaced the voxel values within 4 Å resolution with the low-resolution 3D intensity model. The local update scheme of the EMC algorithm was implemented to reduce the computation time, which limits the rotation samples searched for each data frame to those neighboring the orientations that were given a non-negligible probability in the low-resolution reconstruction [41]. Here the rotations were sampled at order n = 140, which corresponds to an angular resolution of 6.7 mrad. The update was limited to the 3D intensity model, W, because we believe the values of ϕ_k are reliably determined by the low-resolution peaks. The tetragonal and Friedel symmetries were imposed after each update of W to increase the SNR of the Bragg peaks. Figure 5.5 shows the central slice of W perpendicular to the l-axis of the reciprocal lattice, which has the same scale as Figure 5.4(a).

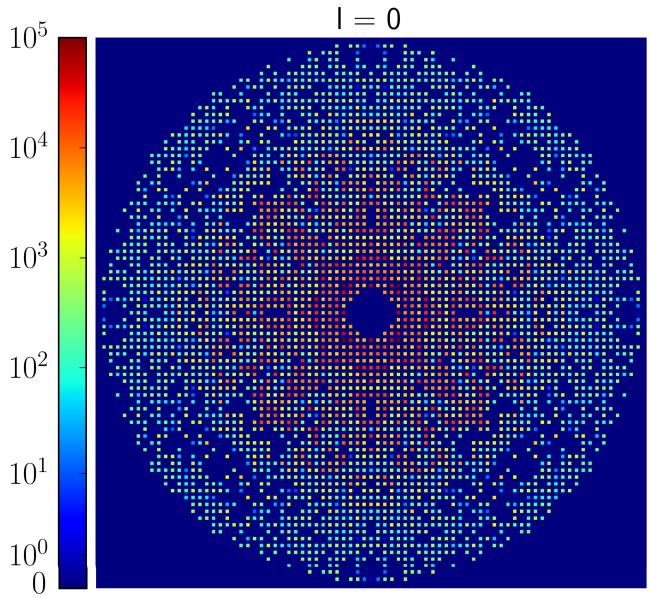

Figure 5.5: The central slice of the high-resolution 3D intensity model, W, perpendicular to the l-axis of the reciprocal lattice, which has the same scale as Figure 5.4(a). The negative reflections were thresholded to zero for rendering.

We evaluated the reproducibility of the reconstruction by $CC_{1/2}$, the correlation coefficient between two sets of Bragg intensities independently reconstructed from two halves of the data frames, respectively. The values of $CC_{1/2}$ were calculated as follows. The 117,904 single-crystal frames were separated into two halves, from which we carried out two independent reconstructions. The reciprocal space was then divided into shells with equal spacing, and the correlation coefficients, $CC_{1/2}$, were computed between the unique reflections from the two reconstructions in each shell. As shown in Figure 5.6, the positive values of $CC_{1/2}$ throughout the spatial frequency magnitudes validate the reproducibility of our approach. The values of $CC_{1/2}$ can further be used to estimate another correlation coefficient, CC^*, through the relation

$$CC^* = \sqrt{\frac{2CC_{1/2}}{1 + CC_{1/2}}}, \tag{5.4}$$

where CC^* measures the correlation between the reconstructed intensities and the underlying true signals [38]. The resolution of the reconstruction is conventionally determined

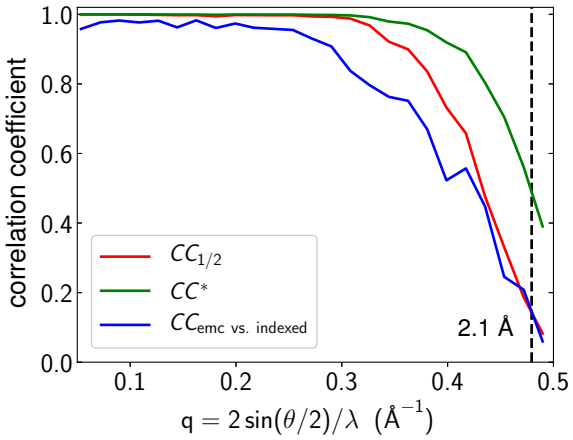

Figure 5.6: The correlation coefficients that validate the quality of our reconstruction. The values of $CC_{1/2}$ show the correlation between Bragg intensities independently reconstructed from two halves of the data frames, respectively. Using Equation (5.4), the values of CC^*, the correlation coefficient between the reconstructed intensities with the underlying true signals, are estimated from the values of $CC_{1/2}$. The other correlation coefficient, $CC_{\text{emc vs. indexed}}$, measures the consistency between our reconstructed intensities with those obtained in Ref. [48] from the indexed frames.

at the value where CC^* drops to 0.5, which corresponds to 2.1 Å in our case.

A more direct validation of our reconstruction comes from the comparison of our reconstructed intensities to those calculated from the indexed peaks using the Mote-Carlo integration approach in Ref. [48]. Dividing the reciprocal space into shells with equal spacing, we calculated the correlation coefficient between the unique reflections from the two sets of Bragg intensities in each shell. Also shown in Figure 5.6, the correlation coefficient stays well above zero up until the resolution cutoff of 2.1 Å, which demonstrates the consistency between the Bragg intensities solved from the two different approaches. When the indexed peaks sufficiently sample crystals of various shapes, sizes and orientations, the Bragg intensities computed by the Monte-Carlo method would in principle correspond to the true signals. In that case, the curve of the correlation coefficient calculated here should move toward the curve of CC^* in Figure 5.6.

5.2.3 Uncertainty estimation

We estimated the uncertainties of the integrated intensities from the measurement, K_{ik}, by error propagation as follows. Let vector $\mathbf{y}$ be a set of functions of vector $\mathbf{x}$. Their covariance matrices, $\Lambda_{\mathbf{y}}$ and $\Lambda_{\mathbf{x}}$, can be related by the formula of error propagation

$$\Lambda_{\mathbf{y}} = J \Lambda_{\mathbf{x}} J^{\top}, \tag{5.5}$$

where J denotes the Jacobian matrix of $\mathbf{y}$. When $\mathbf{x}$ and $\mathbf{y}$ are related by an implicit function, $f(\mathbf{x}, \mathbf{y}) = 0$, the Jacobian matrix is given by

$$J = -\left(\frac{\partial f}{\partial \mathbf{y}}\right)^{-1} \left(\frac{\partial f}{\partial \mathbf{x}}\right). \tag{5.6}$$

From Equation (2.58), the implicit function that relates the photon counts, K_{ik}, and the updated tomogram values, W'_{ij}, is

$$\sum_k P_{jk} \left(p_i \phi_k - \frac{K_{ik}}{b_{ik}/(p_i \phi_k) + W'_{ij}} \right) = 0, \tag{5.7}$$

the derivative of the function to be minimized with respect to W'_{ij}. Since W'_{ij} is a scalar in Equation (5.7), the Jacobian matrix of W'_{ij} becomes a row vector with length N_{data}, the number of data frames, and its k-th element is given by

$$J_k^{ij} = \frac{P_{jk}}{b_{ik}/(p_i \phi_k) + W'_{ij}} \Big/ \sum_{k'} \frac{P_{jk'} K_{ik'}}{(b_{ik'}/(p_i \phi_{k'}) + W'_{ij})^2}. \tag{5.8}$$

The covariance matrix of the measurement, $\Lambda_{\{K_{ik}\}}$, is a diagonal matrix of size $N_{\text{data}} \times N_{\text{data}}$, with the diagonal terms being K_{ik} as a result of the Poisson statistics. Substituting these matrices into Equation (5.5), we obtain the variance of W'_{ij}, denoted by ${\sigma_{W'_{ij}}}^2$.

The values of interest are the uncertainties of the integrated intensities, $I_{hkl} = \sum_{\mathbf{p} \in \{\mathbf{p}_{hkl}\}} W'(\mathbf{p})$, where $\{\mathbf{p}_{hkl}\}$ represents the 3D grid points within the predefined radius, r_p, for the Bragg peak labeled by indices hkl. From Equation (2.52), the variance of

$W'(\mathbf{p})$ is given by

$$\sigma_{W'(\mathbf{p})}{}^2 = \frac{\sum_{ij}\Big[f(\mathbf{p}-\mathbf{R}_j\cdot\mathbf{q}_i)\Big(\sum_k P_{jk}\phi_k\Big)\Big]^2\sigma_{W'_{ij}}{}^2}{\Big[\sum_{ij} f(\mathbf{p}-\mathbf{R}_j\cdot\mathbf{q}_i)\Big(\sum_k P_{jk}\phi_k\Big)\Big]^2}. \tag{5.9}$$

Here we assume that the tomogram values, W'_{ij}, contributing to the same Bragg peak are independent variables. This assumption is based on the observation that each data frame only has non-negligible probabilities at few orientations on convergence, so the values of W'_{ij} with different indices are mostly sampled by different data frames. For the same reason, we also assume that the values, $W(\mathbf{p})$, for $\mathbf{p}$ sampling even the same Bragg peak, are independent variables. The variance of I_{hkl} is hence given by

$$\sigma_{hkl}{}^2 = \sum_{\mathbf{p}\in\{\mathbf{p}_{hkl}\}} \sigma_{W'(\mathbf{p})}{}^2. \tag{5.10}$$

5.3 Structure solution

Model building and refinement steps were done in a manner similar to those performed in Ref. [48], with the intent to validate the EMC approach by a direct comparison to the structure solved from the indexed frames (PDB entry: 5UVJ). The French-Wilson correction [24] was executed to estimate the structure factor magnitudes from the reconstructed weak or negative Bragg intensities. The phases of the structure factors were built from the same template used in Ref. [48] (PDB entry: 4ZIX [25]) using molecular replacement with MOLREP [73]. The structure solution was then iteratively refined to 2.1 Å resolution and inspected using REFMAC5 [50] in the CCP4 suite and Coot [20], respectively. A sodium atom was added as judged by the electron density within the known octahedral coordination of the four residues of the sodium ion. The refinement statistics of the EMC-reconstructed structure solution and the structure solved from the indexed frames, PDB entry: 5UVJ, are summarized in Table 5.1 for comparison. We

	EMC	5UVJ
Resolution (Å)	22.52 – 2.10	35.00 – 2.05
Reflections	7417	7164
Atoms	1019	1023
Protein atoms	1002	1002
Water, ligands and ions	17	21
R_{work}/R_{free} (%)	22.2/28.2	22.8/26.8
R.m.s.d. for bonds (Å)	0.013	0.013
R.m.s.d. for angles (°)	1.211	1.306
Average B value (Å^2)	39.8	34.9
Ramachandran plot statistics (%)		
Favored	96.3	97.6
Allowed	1.3	2.4
Disallowed	0	0
Rotamer outliers	0.93	1

Table 5.1: Refinement statistics of the EMC-reconstructed structure solution and the structure solved from the indexed frames, PDB entry: 5UVJ.

note that the higher average B value of our structure suggests that the data frames we used may have come from less ordered and possibly more weakly diffracting crystals, which are exactly the features we expect from the sparse frames.

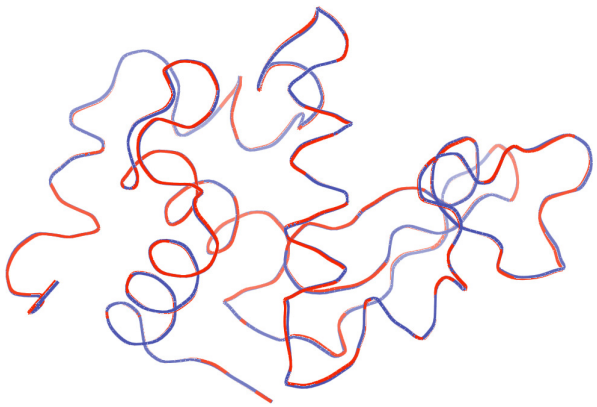

Figure 5.7: Superposition of the ribbon representations of the backbone chains of our structure solution (blue) and the structure, 5UVJ, (red), which presents no significant differences. The C_α atoms between the two structures have r.m.s.d. of 0.131 Å. Deviations greater than this value occur mostly in the solvent-exposed regions, with a maximum deviation of 0.337 Å, though the deviations are only apparent by occasional changes in color from red to blue along the backbone.

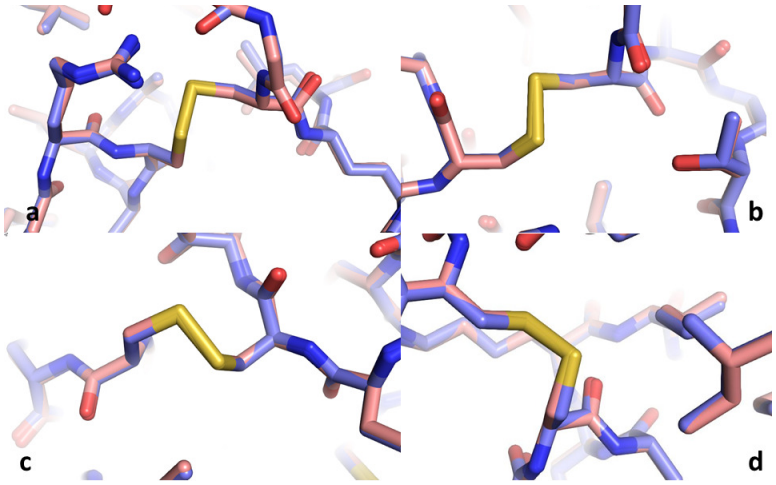

Figure 5.8: Superposition of the four disulfide bonds (yellow) between our structure solution (light red) and the structure 5UVJ (light blue): (a) Cys6-Cys127, (b) Cys30-Cys115, (c) Cys64-Cys80, and (d) Cys76-Cys94. The average deviation for the atoms of the thiol groups is 0.12 Å. Changes are mostly insignificant, and only apparent in splits from light red to light blue.

The structure solved by the EMC approach using the sparse frames compares very well with the structure solved in Ref. [48] using the indexed frames, PDB entry: 5UVJ. Figure 5.7 shows the ribbon representations of the backbone chains of our molecular model (blue) and the structure, 5UVJ, (red). The C_α atoms between the two structures have r.m.s.d. of 0.131 Å, which is visible as an occasional change inbetween the red and blue colors along the backbone chain. Deviations greater than this value occur mostly in the solvent-exposed regions, with the maximum deviation of 0.337 Å. The r.m.s.d. value for the entire protein molecule between the two structures is 0.138 Å, with the maximum deviation of 0.338 Å. Figure 5.8 displays the disulfide bonds (yellow) within two superimposed structures, our structure solution (light red) and 5UVJ (light blue), showing insignificant deviations between the structures in the more radiation damage-prone bonds. The average deviation for the atoms of the thiol groups is 0.12 Å.

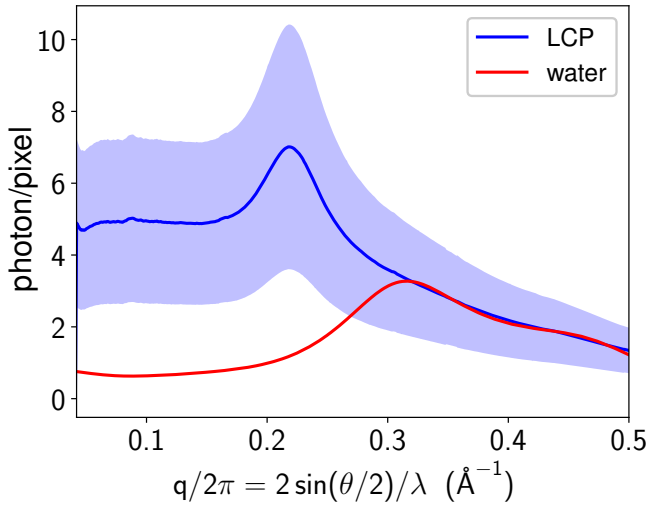

Figure 5.9: The scattering profiles of LCP and water, which were generated by the weighted average of the azimuthally symmetric background estimates for each frame and simulation, respectively. The shaded region is within one standard deviation from the average scattering profile of LCP. The large standard deviation is mainly caused by the jittering of the LCP stream.

5.4 Discussion

The major source of error that limits the quality of our reconstruction is the high background scatter from LCP. From the estimated X-ray beam size (different beam sizes of 5, 10 or 20 μm were used at different times during the data collection), the diameter of the LCP gel column (50 μm), and the reconstructed crystal volumes (Figure 5.4(b)), we can estimate the total number of photons scattered by LCP to be tens to thousands times more than that scattered by the crystal in each data frame. As a result, the weak crystal intensities are substantially affected by background intensity fluctuations.

The high background scattering from LCP is best shown when compared with the scattering profile of water. Assuming an X-ray beam size of 10 μm and a detector exposure time of 0.1 second, we simulated the scattering profile from a water column of 50 μm diameter from the experimentally measured pair distribution function [52, 67]. The scattering profile of LCP was obtained by the average of the azimuthally symmetric

background estimates for each data frame, which were rescaled to have the same nominal X-ray beam size and detector exposure time before the average. As shown in Figure 5.9, LCP scatters a large number of photons within 3 Å resolution, and this has motivated search for sample delivery media that scatter fewer background photons. For example, agarose was used in Ref. [12] to reduce background scattering, although the agarose stream tends to be unstable under ambient pressure. On the other hand, the sodium carboxymethyl cellulose (NaCMC) and poly(ethylene oxide) (PEO) reported in Ref. [40] and [48], respectively, produce stable streams and lower background scattering than LCP, and therefore may be good substitutes for LCP. Another option for background reduction is to use the fixed-target approach. As recently demonstrated in Ref. [57] and [62], rapid data collection can be achieved by fast scanning through micro-patterned silicon chips mounted with protein microcrystals. Nevertheless, the challenge of the chip methods is to avoid preferential crystal orientations. Other possible methods include microcrystal droplets deposited on low-background tape carriers [26].

Our EMC-based analysis method provides a means to make use of the crystal diffraction patterns whose signals are too noisy to be considered by the prior state-of-the-art. In particular, the weak crystal diffraction signals can be extracted from the diffuse background scattering to obtain the Bragg intensities. This approach reduces the sample consumption by making use of all the available data frames. The reconstruction of the crystal volume distribution may also be useful for the development of the sample injection technology. As shown in the proof-of-concept studies in Chapter 4, reconstruction is feasible for crystal sizes as small as 1 to 2 μm within tolerable radiation dose if the background scatter can be sufficiently reduced. The successful application of our approach to SMX data collected from such small crystals will be a great advance in protein structure determination at storage ring sources, and at the same time ease the high demands for XFEL beamtime.

CHAPTER 6

CONCLUSIONS

As single-particle cryoEM has become a competitive high-resolution technique for structure determination, X-ray methods have started to develop a niche in probing the dynamics of biological macromolecules by external perturbations [66, 68]. In order to rapidly excite structural changes, the probed samples have to be small in size, for example, individual particles or microcrystals. This not only requires constant improvements in experimental technology to measure the weak signals scattered from the small samples, but also advanced analysis tools to extract useful information from the noise-limited signals. Our work in this thesis serves as a timely contribution to the latter need.

The goal of this thesis is to give a theoretical overview on the development of the EMC algorithm and its applications in structural biology using X-ray methods. The basic principles behind X-ray diffraction measurements are described primarily in conventions adopted by physicists, with the hope to better explain the physical meaning of the commonly used quantities by practitioners. We also categorize the variants of the EMC algorithm to allow a systematic choice of appropriate models in view of different experimental conditions.

In this thesis we have discussed two main applications of the EMC algorithm — SPI and storage-ring based SMX. The development of SPI is currently limited by the lack of data. The rate that individual particles are intercepted by X-ray pulses is currently insufficient to allow high-resolution 3D reconstructions. Moreover, the quality of reconstructions is degraded by structural heterogeneity from the adsorption of non-volatile contaminants on the particles. If these issues can be resolved, for example by advances in injector technology, SPI could become an unparalleled tool to study the dynamics of isolated macromolecules with time resolutions up to femtoseconds. Using the EMC

algorithm, we have demonstrated that protein structure determination is feasible from unindexable data frames collected from HEWL microcrystals. Once this analysis approach is shown to be applicable to weakly scattering microcrystals, such as those formed by membrane proteins, it will make SMX an attractive approach for protein structure determination because of the wide availability of beamtime at storage ring sources. Continued development of lower-background microcrystal carrier methods will facilitate the application of our method.

APPENDIX A

TUTORIAL ON CRYSTAL INTENSITY RECONSTRUCTION

This appendix gives instruction on reconstructing Bragg intensities from SMX data using the EMC algorithm[1]. The default input data format for our program is *cbf*. We will demonstrate the step-by-step analysis using a subset of the HEWL crystal diffraction patterns collected by Martin-Garcia et al. at the Advanced Photon Source [48]. The workflow of the analysis is adapted from that described in Chapter 5, and is illustrated in Figure A.1. The program was written in *C* and *Python*, and is executed on *Linux* using the *MPI* parallelization framework. The required packages are

- Requirements for *C*: *gcc*, *OpenMPI* and *OpenSSL*.
- Requirements for *Python*: *Python2.7*, *NumPy*, *Matplotlib* and *FabIO*.
- *Git*, *X Window System*.

A.1 Initialization

Here we download the data and the source code of the analysis software, and set up the environment for data processing. The data is deposited on the Coherent X-ray Imaging Data Bank (CXIDB) [47] and can be downloaded from the website

`http://cxidb.org/data/82/raw-data`

to the hard drive of a local computer cluster. The source code can be obtained by executing the command

`git clone git@github.com:tl578/EMC-for-SMX.git,`

and this creates `EMC-for-SMX`, which is the working directory for the intensity reconstruction. The relevant files and modules in the working directory include:

[1]The up-to-date tutorial can also be found at `https://github.com/tl578/EMC-for-SMX/wiki`.

1. `config.ini`: configuration file that records the reconstruction parameters.
2. `init-recon.py`: python script that initializes the reconstruction.
3. `aux`: directory that stores auxiliary files.
4. `make-detector`: maps detector pixels to reciprocal space.
5. `make-background`: generates pixel-wise background estimates and identifies Bragg peak candidates.
6. `make-powder`: generates pseudo-powder patterns.
7. `make-quaternion`: generates rotation samples.
8. `orient-peak`: finds probable orientations for each data frame.
9. `reduce-data`: converts data to the format used by the EMC algorithm.
10. `make-Ematrix`: creates mapping between Bragg peaks and detector pixels at different orientations.
11. `low-res-emc`: low-resolution reconstruction using the standard update scheme of the EMC algorithm.
12. `rej-frames`: rejects frames with no or multiple crystals.
13. `setup-local`: creates the necessary files for the high-resolution reconstruction.
14. `local-update`: high-resolution reconstruction using the local update scheme of the EMC algorithm.
15. `cal-CC`: splits data into two halves to calculate the correlation coefficient, CC^*.

The initialization step is completed by executing the command

`python init-recon.py [reduced-data-dir],`

where `[reduced-data-dir]` is the path to the directory that will be used to store the reduced data.

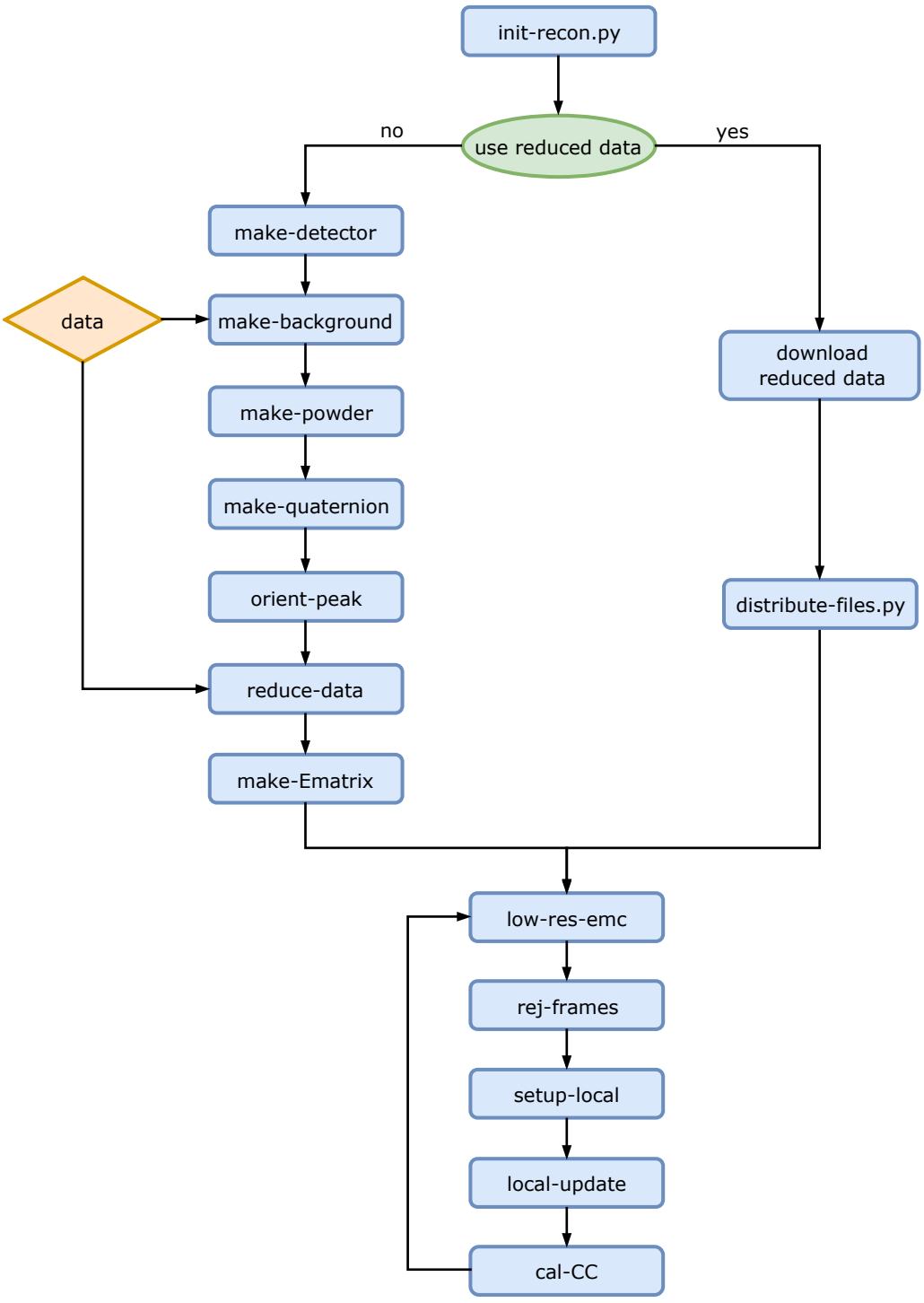

Figure A.1: Flowchart of the analysis of SMX data using our software package.

A.2 Data reduction

In this section, we generate the necessary files for the EMC reconstruction. This step can be skipped by directly using the reduced data deposited on CXIDB for the HEWL dataset, which will be explained in Section A.2.7.

A.2.1 Mapping detector pixels

We start the analysis by filling in the experimental parameters in the configuration file, `config.ini`. The parameters in the [`make-detector`] section of `config.ini` are:

```
[make-detector]
# pixel
num_row = 2527
num_col = 2463
cx = 1285.5
cy = 1262.0
Rstop = 115.0

# meter
detd = 0.45
px = 172e-6

# angstrom
wl = 1.03324
res_max = 1.95

# beam incidence direction
sx = 0.005
sy = -0.01
sz = -1.0
```

The parameters `num_row` and `num_col` specify the detector size in pixels. The pixels are labeled by coordinates $(\mathtt{x}, \mathtt{y})$, with $\mathtt{x} = 0, 1, \ldots, \mathtt{num\_row} - 1$ and $\mathtt{y} = 0, 1, \ldots, \mathtt{num\_col} - 1$. With this choice of coordinates, the upper-left detector pixel

has coordinates (0, 0), and the X-ray beam is incident in the $-\hat{z}$ direction. Since the main application of our program is the analysis of SMX data collected at storage ring sources, the X-ray beam polarization is assumed to be in the $\hat{y}$ direction. The parameters (cx, cy) label the beam incidence point on the detector, and Rstop is the beamstop radius in pixels. The other parameters include detd, the sample-to-detector distance, px, the squared detector pixel size, wl, the incident X-ray wavelength, and res_max, the maximum full-period resolution of the pixels that will be considered in the reconstruction. Finally, the vector (sx, sy, sz) indicates the beam incidence direction (does not have to be normalized), and is typically set as (0, 0, −1).

We then move to the directory make-detector, and execute the command

python make-mask.py [path to frame] > run.log

to create the file mask.dat in the directory aux to mask out the detector gaps and the pixels shadowed by the beamstop holder, where [path to frame] is the path to one of the data frames in the *cbf* format. The files that record the mapping of the detector pixels to reciprocal space are obtained with the commands

gcc make-detector.c -O3 -lm -o det

./det ../config.ini >> run.log.

A.2.2 Background estimation and peak finding

After moving to the directory make-background, we generate the lists of the filenames associated with each data frame using the command

python make-filelists.py [raw-data-dir],

where [raw-data-dir] is the path to the directory that contains the *cbf* files downloaded from CXIDB. Subsequently, we update the parameters in the [make-background] section in config.ini:

```
[make-background]
num_raw_data = 79992
hot_pix_thres = 1e4
qlen = 500
```

The execution of the command above has automatically updated the value of num_raw_data, the total number of data frames. The parameter hot_pix_thres is the threshold value beyond which a pixel is identified as defective and masked out. In our analysis, we assume that the background scatter in each data frame is azimuthally symmetric about the incident X-ray beam, and qlen represents the number of bins that divide the spatial frequency magnitudes with equal spacing for the background estimation. Finally, we execute the commands

```
make
mpirun -np [nproc] ./ave_bg ../config.ini > run.log &
```

to estimate the pixel-wise background values and identify the outlier pixels in each frame, where [nproc] is the number of processors used in the parallel processing.

A.2.3 Lattice parameter estimation

Next, we move to the directory make-powder to estimate the lattice parameters. The parameters in the [make-powder] section in config.ini are:

```
[make-powder]
min_patch_sz = 2
max_patch_sz = 10
```

```
min_num_peak = 3
max_num_peak = 20
```

A Bragg peak candidate is assumed to contain at least `min_patch_sz` but no more than `max_patch_sz` contiguous outlier pixels identified from the diffuse background scatter. Only the data frames with at least `min_num_peak` but no more than `max_num_peak` candidate peaks are kept for the later analysis. The enforcement of data sparsity can be removed by making `max_num_peak` a large integer.

By executing the commands

```
gcc make-powder.c -O3 -lm -o powder
./powder ../config.ini > run.log,
```

we generate the files `frame-peak-count.dat`, `patch-sz-count.dat`, `1d-pseudo-powder.dat` and `2d-pseudo-powder.dat`. The number of candidate peaks in each data frame is recorded in `frame-peak-count.dat` The file `patch-sz-count.dat` represents the histogram of the size of contiguous outlier pixels, from which we can check if the original choice of `max_patch_sz` is reasonable. The file `1d-pseudo-powder.dat` contains three columns: the spatial frequency magnitudes, the counts of inter-peak distances in reciprocal space in each frame, and the counts of spatial frequency magnitudes of the candidate peaks. Finally, the file `2d-pseudo-powder.dat` records the maximum photon count in each detector pixel.

In the analysis of the test dataset, we fit the lattice parameters $a = 79.1$ Å and $c = 38.4$ Å by assuming a primitive tetragonal lattice. This choice can be assessed by executing the command

```
python plot-1d-powder.py
```

to plot the histograms of the inter-peak distances and the spatial frequency magnitudes of the candidate peaks. For general crystal lattices, the lattice parameters have to be

estimated by fitting the histogram of the inter-peak distances. By executing the command

`python plot-2d-powder.py,`

we plot the 2D pseudo-powder pattern to check if the original estimates of the parameters (`cx`, `cy`), the beam incidence point on the detector, and (`sx`, `sy`, `sz`), the beam incidence direction, are reasonable. The whole data processing from Section A.2.1 to here should be rerun if these parameters have to be changed. The values of the estimated lattice parameters are stored as a 3×3 matrix

```
u[0] v[0] w[0]
u[1] v[1] w[1]
u[2] v[2] w[2]
```

in the file `basis-vec.dat` in the directory `aux`, where $\vec{u}$, $\vec{v}$ and $\vec{w}$ denote the basis vectors of the primitive unit cell in units of Å. This file should be created by the user for general crystal lattices.

A.2.4 Finding probable orientations

Our next step is to narrow down the number of probable orientations for each frame by directly rotating the centroids of the Bragg peak candidates over all rotation samples in reciprocal space — an orientation is kept for a particular frame if at least `min_num_peak` candidate peaks overlap with the predicted Bragg peaks. We begin by choosing the parameters in the [`orient-peak`] section of `config.ini`:

```
[orient-peak]
res_cutoff = 4.0
VN = 15
gw = 2.0
```

The parameter res_cutoff specifies the highest full-period resolution of data in unit of Å that will be used in determining the probable orientations and the low-resolution EMC reconstruction. The parameter VN denotes the number of voxels between the closest Bragg peaks in reciprocal space, and gw is the radius of a Bragg peak in unit of voxel.

After updating the parameters, we move to the directory make-quaternion to generate the rotation samples with the command

python make-rot-samples.py [num_div].

The integer [num_div] specifies the angular resolution $\delta\theta = 0.944/[\texttt{num\_div}]$. An angular resolution of at least $(2\texttt{gw} \cdot \texttt{res\_cutoff} \cdot \texttt{min\_rcell})/\texttt{VN}$ is required in order to not to miss any Bragg peaks. Here min_rcell denotes the minimum peak distance in reciprocal space, with unit of Å^{-1}. For the test dataset, we have $\texttt{min\_rcell} = 1/a$, and the resulting angular resolution corresponds to [num_div] = 70. This command creates the file c-quaternion[num_div].bin in the directory aux. Finally, we move to the directory orient-peak, and execute the commands

mpicc mpi-sync-orient-peak.c -O3 -lm -o orient

mpirun -np [nproc] ./orient ../config.ini > run.log &

to find the probable orientations for each frame, where [nproc] is the number of processors to be used. The output file num_prob_orien.dat records the number of probable orientations for each data frame.

A.2.5 Data conversion

Subsequently, we move to the directory reduce-data to convert data to the format that will be used by the EMC reconstruction. In the [reduce-data] section of config.ini, we have the parameters:

```
[orient-peak]
nproc = 20
mpi_bgfile = [reduced-data-dir]/Data/mpi-bg_model.bin
mpi_datafile = [reduced-data-dir]/Data/mpi-datafile.bin
```

Here `nproc` is the number of processors that will be used for the EMC reconstruction, and `[reduced-data-dir]` is the directory we used in Section A.1. The files `mpi_bgfile` and `mpi_datafile` store the background estimates and photon counts of the data frames that will be input to the EMC algorithm, respectively. In order to reduce the time spent on reading data, the photon counts are stored as short integers. The frames with more than `max_num_peak` identified peaks or no probable orientations found in Section A.2.4 will be excluded.

We generate `mpi_bgfile` and other auxiliary files by executing the commands

```
gcc reduce-data.c -O3 -lm -o reduce-data
./reduce-data ../config.ini > run.log.
```

The order of the frames is rearranged to balance the work loads between the `nproc` processors based on the number of probable orientations per frame, and this information is stored in the file `reduced-data_id.dat`. The file `mpi_datafile` is generated using the commands

```
make
./wr-data ../config.ini >> run.log &.
```

A.2.6 Expansion matrix calculation

Since the crystal diffraction signals are concentrated in the Bragg spots, we can speed up the expand (E) step of the EMC reconstruction by precalculating a look-up table that

records the mapping between the Bragg peaks and the detector pixels over all rotation samples. This can be done by moving to the directory `make-Ematrix` and executing the commands

`mpicc mpi-make-Emat.c -O3 -lm -o emat`

`mpirun -np [nproc] ./emat -low ../config.ini > run.log &,`

where [`nproc`] is the number of processors to be used. The mapping is stored in the files `r2peak_file` and `peak2r_file`, whose locations are specified in the [`make-Ematrix`] section of `config.ini`.

A.2.7 Skipping data reduction

For those who would like to try an EMC reconstruction immediately, we have provided the reduced data generated from the HEWL dataset following the data processing procedures described above. After creating the directory `skip-data-reduction`, we can download the reduced data from the website

`http://cxidb.org/data/82/reduced-data.`

The downloaded files are moved to their appropriate locations by executing the command

`python distribute-files.py [work-dir] > dist.log &`

to get ready for the EMC reconstruction. Here [`work-dir`] denotes the path to the working directory, `EMC-for-SMX`.

A.3 Low-resolution EMC reconstruction

Now we proceed with the low-resolution intensity reconstruction using the standard update scheme of the EMC algorithm. We first update the parameters in the [low-res-emc] section of config.ini:

```
[low-res-emc]
iter_data_block = 5
prob_dir = [reduced-data-dir]/Data/high-prob
prob_orien_file = [work-dir]/aux/prob-orien.bin
reduced_data_id_file = [work-dir]/reduce-data/reduced-data_id.dat
start_phi_file = [work-dir]/aux/start-phi.dat
start_intens_file = [work-dir]/aux/start_intensity.bin
```

Since the data size is generally several hundred GB or more, the data frames are separated into iter_data_block blocks and read in sequentially in each EMC iteration to save memory. The directory prob_dir stores the output files of the reconstruction. The file prob_orien_file records the probable orientations for each data frame, and reduced_data_id_file stores the original data frame order in reduce-data, before the rearrangement. The files start_phi_file and start_intens_file are the initial models for the reconstruction.

Next, we move to the directory aux to generate start_phi_file and sym-op.dat, which stores the symmetry operators of the crystal lattice. The operations in this paragraph can be skipped if the reduced data downloaded from CXIDB is used. By executing the command

python init-phi.py,

the initial values of the scale factors, ϕ_k, are estimated with the average peak value in each data frame. Two additional files start-phi-A.dat and start-phi-B.dat are also generated by this command. These files store the scale factors for the two independent halves of the data frames, which will be used in Section A.5. The file sym-op.dat is

generated with the command

```
python make-sym-op.py
```

for the tetragonal crystal lattice. This file should be created by the user for general crystal lattices.

The EMC reconstruction is started by executing the commands

```
make
```

```
mpirun -np [nproc] ./emc ../config.ini [iter] > run.log &,
```

where [nproc] should be the same as the value specified in config.ini, and [iter] is the number of EMC iterations. The 3D intensity model is initialized by placing 3D Gaussian of random height at each Bragg position, whose values are stored in start_intens_file. In the n^{th} EMC iteration, our program creates two directories in prob_dir: iter_flag-[2n − 1] and iter_flag-[2n], which store the outputs from the updates of the intensity model and the scale factors, ϕ_k, respectively. In order to resume a previous reconstruction, the user has to replace the files, start_phi_file and start_intens_file, by total-phi.dat and finish_intensity.bin output in the last iteration of the previous reconstruction. The output files from the previous reconstruction have to be moved elsewhere to avoid being overwritten.

After the reconstruction reaches convergence, we execute the command

```
python move-recon-files.py ../config.ini
```

to create the directory low-res-recon in prob_dir, and move start_phi_file, start_intens_file and the output files of the reconstruction there. Finally, we move to the directory rej-frames and execute the command

```
python rej-frames.py ../config.ini
```

to exclude the frames that contain no or multiple crystals. This command creates an updated `start_phi_file`, where the excluded frames will have the scale factors, ϕ_k, set as zero.

A.4 High-resolution EMC reconstruction

Here we implement the local update scheme of the EMC algorithm to extend the reconstruction to high resolution based on the converged models and probability distribution given by the low-resolution reconstruction. We first choose the resolution cutoff that will be used in the high-resolution EMC reconstruction. This value is specified by the parameter `high_res_cutoff` in `config.ini`, and should be larger than `res_max` but smaller than `res_cutoff`.

Next, we move to the directory `setup-local`, and execute the command

`python setup-quat.py ../config.ini >> run.log &`

to generate the rotation samples that will be used and the file that stores the mapping between this rotation sampling and that used in the low-resolution reconstruction. The angular resolution of the new rotation samples is chosen to not to miss any Bragg peaks within the resolution `high_res_cutoff`. By executing the command

`python setup-intens.py ../config.ini >> run.log &,`

we generate the initial 3D intensity model for the high-resolution reconstruction, which is stored in the file `start_intens_file`. Finally, the mapping between Bragg peaks and detector pixels for the new rotation samples is generated using the command

`python setup-Ematrix.py ../config.ini >> run.log &.`

This mapping is stored in the files `local_r2peak_file` and `local_peak2r_file`, as specified in `config.ini`.

After moving to the directory `local-update`, we start the high-resolution reconstruction using the commands

```
make
```

```
mpirun -np [nproc] ./emc ../config.ini [iter] > run.log &.
```

The parameter [nproc] should be the same as the value specified in `config.ini`, and [iter] is the number of EMC iterations. The output files of the n^{th} EMC iteration are stored in the directory `prob_dir/iter_flag-[n]`. When the reconstruction reaches convergence, we execute the command

```
python move-recon-files.py ../config.ini
```

to create the directory `high-res-recon` in `prob_dir`, and move `start_phi_file`, `start_intens_file` and the output files of the reconstruction there.

A.5 Resolution estimation

We estimate the resolution of the reconstruction by calculating the correlation coefficient, CC^*, whose value can be estimated from another correlation coefficient, $CC_{1/2}$, through Equation (5.4). Moving to the directory, `cal-CC`, we execute the command

```
python split-data.py > run.log &
```

to separate the data frames into two halves and generate the corresponding two configuration files, `config-A.ini` and `config-B.ini`, in the working directory. Independent reconstructions using the two halves of the data frames are completed by repeating the

procedures in Sections A.3 and A.4, with the argument, `config.ini`, replaced by either `config-A.ini` or `config-B.ini`. After completing the two independent reconstructions, we move back to the directory, `cal-CC`, and calculate the correlation coefficient, CC^*, by executing the command

`python cal-CC.py.`

The resolution is conventionally determined by the spatial frequency magnitude where CC^* drops to 0.5.

BIBLIOGRAPHY

[1] A. Aquila, A. Barty, C. Bostedt, S. Boutet, G. Carini, D. dePonte, P. Drell, S. Doniach, K. H. Downing, T. Earnest, H. Elmlund, V. Elser, M. GÃijhr, J. Hajdu, J. Hastings, S. P. Hau-Riege, Z. Huang, E. E. Lattman, F. R. N. C. Maia, S. Marchesini, A. Ourmazd, C. Pellegrini, R. Santra, I. Schlichting, C. Schroer, J. C. H. Spence, I. A. Vartanyants, S. Wakatsuki, W. I. Weis, and G. J. Williams. The linac coherent light source single particle imaging road map. *Structural Dynamics*, 2(4):041701, 2015.

[2] K.-J. Armache, S. Mitterweger, A. Meinhart, and P. Cramer. Structures of Complete RNA Polymerase II and its Subcomplex, Rpb4/7. *The Journal of Biological Chemistry*, 280(8):7131–7134, 2005.

[3] K. Ayyer, T.-Y. Lan, V. Elser, and N. D. Loh. *Dragonfly*: an implementation of the expand–maximize–compress algorithm for single-particle imaging. *Journal of Applied Crystallography*, 49(4):1320–1335, 2016.

[4] K. Ayyer, H. T. Philipp, M. W. Tate, V. Elser, and S. M. Gruner. Real-Space x-ray tomographic reconstruction of randomly oriented objects with sparse data frames. *Opt. Express*, 22(3):2403–2413, 2014.

[5] K. Ayyer, H. T. Philipp, M. W. Tate, J. L. Wierman, V. Elser, and S. M. Gruner. Determination of crystallographic intensities from sparse data. *IUCrJ*, 2(1):29–34, 2015.

[6] X. Bai, T. G. Martin, S. H. W. Scheres, and H. Dietz. Cryo-EM structure of a 3D DNA-origami object. *Proceedings of the National Academy of Sciences*, 109(49):20012–20017, 2012.

[7] A. Barty, R. A. Kirian, F. R. N. C. Maia, M. Hantke, C. H. Yoon, T. A. White, and H. N. Chapman. *Cheetah*: software for high-throughput reduction and analysis of serial femtosecond X-ray diffraction data. *Journal of Applied Crystallography*, 47(3):1118–1131, 2014.

[8] B. Bhyravbhatla, S. J. Watowich, and D. L. Caspar. Refined atomic model of the four-layer aggregate of the tobacco mosaic virus coat protein at 2.4-A resolution. *Biophysical Journal*, 74(1):604–615, 1998.

[9] S. Botha, K. Nass, T. R. M. Barends, W. Kabsch, B. Latz, F. Dworkowski, L. Foucar, E. Panepucci, M. Wang, R. L. Shoeman, I. Schlichting, and R. B. Doak. Room-temperature serial crystallography at synchrotron X-ray sources using slowly flow-

ing free-standing high-viscosity microstreams. *Acta Crystallographica Section D*, 71(2):387–397, 2015.

[10] S. Boutet, L. Lomb, G. J. Williams, T. R. M. Barends, A. Aquila, R. B. Doak, U. Weierstall, D. P. DePonte, J. Steinbrener, R. L. Shoeman, M. Messerschmidt, A. Barty, T. A. White, S. Kassemeyer, R. A. Kirian, M. M. Seibert, P. A. Montanez, C. Kenney, R. Herbst, P. Hart, J. Pines, G. Haller, S. M. Gruner, H. T. Philipp, M. W. Tate, M. Hromalik, L. J. Koerner, N. van Bakel, J. Morse, W. Ghonsalves, D. Arnlund, M. J. Bogan, C. Caleman, R. Fromme, C. Y. Hampton, M. S. Hunter, L. C. Johansson, G. Katona, C. Kupitz, M. Liang, A. V. Martin, K. Nass, L. Redecke, F. Stellato, N. Timneanu, D. Wang, N. A. Zatsepin, D. Schafer, J. Defever, R. Neutze, P. Fromme, J. C. H. Spence, H. N. Chapman, and I. Schlichting. High-resolution protein structure determination by serial femtosecond crystallography. *Science*, 337(6092):362–364, 2012.

[11] H.N. Chapman, P. Fromme, A. Barty, T.A. White, R.A. Kirian, A. Aquila, M.S. Hunter, J. Schulz, D.P. DePonte, U. Weierstall, R.B. Doak, F.R.N.C. Maia, A.V. Martin, I. Schlichting, L. Lomb, N. Coppola, R.L. Shoeman, S.W. Epp, R. Hartmann, D. Rolles, A. Rudenko, L. Foucar, N. Kimmel, G. Weidenspointner, P. Holl, M. Liang, M. Barthelmess, C. Caleman, S. Boutet, M.J. Bogan, J. Krzywinski, C. Bostedt, S. Bajt, L. Gumprecht, B. Rudek, B. Erk, C. Schmidt, A. HÃűmke, C. Reich, D. Pietschner, L. StrÃijder, G. Hauser, H. Gorke, J. Ullrich, S. Herrmann, G. Schaller, F. Schopper, H. Soltau, K.-U. KÃijhnel, M. Messerschmidt, J.D. Bozek, S.P. Hau-Riege, M. Frank, C.Y. Hampton, R.G. Sierra, D. Starodub, G.J. Williams, J. Hajdu, N. Timneanu, M.M. Seibert, J. Andreasson, A. Rocker, O. JÃűnsson, M. Svenda, S. Stern, K. Nass, R. Andritschke, C.-D. SchrÃűter, F. Krasniqi, M. Bott, K.E. Schmidt, X. Wang, I. Grotjohann, J.M. Holton, T.R.M. Barends, R. Neutze, S. Marchesini, R. Fromme, S. Schorb, D. Rupp, M. Adolph, T. Gorkhover, I. Andersson, H. Hirsemann, G. Potdevin, H. Graafsma, B. Nilsson, and J.C.H. Spence. Femtosecond X-ray protein nanocrystallography. *Nature*, 470(7332):73–77, 2011.

[12] C. E. Conrad, S. Basu, D. James, D. Wang, A. Schaffer, S. Roy-Chowdhury, N. A. Zatsepin, A. Aquila, J. Coe, C. Gati, M. S. Hunter, J. E. Koglin, C. Kupitz, G. Nelson, G. Subramanian, T. A. White, Y. Zhao, J. Zook, S. Boutet, V. Cherezov, J. C. H. Spence, R. Fromme, U. Weierstall, and P. Fromme. A novel inert crystal delivery medium for serial femtosecond crystallography. *IUCrJ*, 2(4):421–430, 2015.

[13] B. J. Daurer, M. F. Hantke, C. Nettelblad, and F. R. N. C. Maia. *Hummingbird*: monitoring and analyzing flash X-ray imaging experiments in real time. *Journal of Applied Crystallography*, 49(3):1042–1047, 2016.

[14] B. J. Daurer, K. Okamoto, J. Bielecki, F. R. N. C. Maia, K. Mühlig, M. M. Seibert, M. F. Hantke, C. Nettelblad, W. H. Benner, M. Svenda, N. Tîmneanu, T. Ekeberg, N. D. Loh, A. Pietrini, A. Zani, A. D. Rath, D. Westphal, R. A. Kirian, S. Awel, M. O. Wiedorn, G. van der Schot, G. H. Carlsson, D. Hasse, J. A. Sellberg, A. Barty, J. Andreasson, S. Boutet, G. J. Williams, J. Koglin, I. Andersson, J. Hajdu, and D. S. D. Larsson. Experimental strategies for imaging bioparticles with femtosecond hard X-ray pulses. *IUCrJ*, 4(3):251–262, 2017.

[15] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B.*, 39(1):1–38, 1976.

[16] T. Ekeberg, M. Svenda, C. Abergel, F. R. N. C. Maia, V. Seltzer, J.-M. Claverie, M. Hantke, O. Jönsson, C. Nettelblad, G. van der Schot, M. Liang, D. P. DePonte, A. Barty, M. M. Seibert, B. Iwan, I. Andersson, N. D. Loh, A. V. Martin, H. N. Chapman, C. Bostedt, J. D. Bozek, K. R. Ferguson, J. Krzywinski, S. W. Epp, D. Rolles, A. Rudenko, R. Hartmann, N. Kimmel, and J. Hajdu. Three-Dimensional Reconstruction of the Giant Mimivirus Particle with an X-Ray Free-Electron Laser. *Physical Review Letter*, 114:098102, 2015.

[17] V. Elser. Phase retrieval by iterated projections. *Journal of the Optical Society of America A*, 20(1):40–55, 2003.

[18] V. Elser, T.-Y. Lan, and T. Bendory. Benchmark problems for phase retrieval. *ArXiv e-prints*, 2017.

[19] V. Elser and R. P. Millane. Reconstruction of an object from its symmetry-averaged diffraction pattern. *Acta Crystallographica Section A*, 64(2):273–279, 2008.

[20] P. Emsley, B. Lohkamp, W. G. Scott, and K. Cowtan. Features and development of *Coot*. *Acta Crystallographica Section D*, 66(4):486–501, 2010.

[21] K. R. Ferguson, M. Bucher, J. D. Bozek, S. Carron, J.-C. Castagna, R. Coffee, G. I. Curiel, M. Holmes, J. Krzywinski, M. Messerschmidt, M. Minitti, A. Mitra, S. Moeller, P. Noonan, T. Osipov, S. Schorb, M. Swiggers, A. Wallace, J. Yin, and C. Bostedt. The Atomic, Molecular and Optical Science instrument at the Linac Coherent Light Source. *Journal of Synchrotron Radiation*, 22(3):492–497, 2015.

[22] J. R. Fienup. Phase retrieval algorithms: a comparison. *Applied optics*, 21(15):2758–2769, 1982.

[23] J. Frank. *Three-Dimensional Electron Microscopy of Macromolecular Assemblies*. Oxford University Press, 2nd edition, 2006.

[24] S. French and K. Wilson. On the treatment of negative intensity observations. *Acta Crystallographica Section A*, 34(4):517–525, 1978.

[25] R. Fromme, A. Ishchenko, M. Metz, S. R. Chowdhury, S. Basu, S. Boutet, P. Fromme, T. A. White, A. Barty, J. C. H. Spence, U. Weierstall, W. Liu, and V. Cherezov. Serial femtosecond crystallography of soluble proteins in lipidic cubic phase. *IUCrJ*, 2(5):545–551, 2015.

[26] F. D Fuller, S. Gul, R. Chatterjee, E. S. Burgie, I. D. Young, H. Lebrette, V. Srinivas, A. S. Brewster, T. Michels-Clark, J. A. Clinger, B. Andi, M. Ibrahim, E. Pastor, C. de Lichtenberg, R. Hussein, C. J. Pollock, M. Zhang, C. A. Stan, T. Kroll, T. Fransson, C. Weninger, M. Kubin, P. Aller, L. Lassalle, P. BrÃđuer, M. D. Miller, M. Amin, S. Koroidov, C. G. Roessler, M. Allaire, R. G. Sierra, P. T. Docker, J. M. Glownia, S. Nelson, J. E. Koglin, D. Zhu, M. Chollet, S. Song, H. Lemke, M. Liang, D. Sokaras, R. Alonso-Mori, A. Zouni, J. Messinger, U. Bergmann, A. K. Boal, J. M. Bollinger Jr, C. Krebs, M. HÃűgbom, G. N. Phillips Jr, R. D. Vierstra, N. K. Sauter, A. M. Orville, J. Kern, V. K. Yachandra, and J. Yano. *Nature Methods*, 14:443–449, 2017.

[27] C. Gati, G. Bourenkov, M. Klinge, D. Rehders, F. Stellato, D. Oberthür, O. Yefanov, B. P. Sommer, S. Mogk, M. Duszenko, C. Betzel, T. R. Schneider, H. N. Chapman, and L. Redecke. Serial crystallography on *in vivo* grown microcrystals using synchrotron radiation. *IUCrJ*, 1(2):87–94, 2014.

[28] C. Gatsogiannis and J. Markl. Keyhole Limpet Hemocyanin: 9 Å CryoEM Structure and Molecular Model of the KLH1 Didecamer Reveal the Interfaces and Intricate Topology of the 160 Functional Units. *Journal of Molecular Biology*, 385(3):963–983, 2009.

[29] S. M. Gruner and E. E. Lattman. Biostructural Science Inspired by Next-Generation X-Ray Sources. *Annual Review of Biophysics*, 44(1):33 – 51, 2015.

[30] T. Hahn, editor. *International tables for crystallography*, volume A. Chester: International Union of Crystallography, 1st online edition, 2006.

[31] B.L. Henke, E.M. Gullikson, and J.C. Davis. X-Ray Interactions: Photoabsorption, Scattering, Transmission, and Reflection at E = 50-30,000 eV, Z = 1-92. *Atomic Data and Nuclear Data Tables*, 54(2):181–342, 1993.

[32] M. Heymann, A. Opthalage, J. L. Wierman, S. Akella, D. M. E. Szebenyi, S. M. Gruner, and S. Fraden. Room-temperature serial crystallography using a kinetically optimized microfluidic device for protein crystallization and on-chip X-ray diffraction. *IUCrJ*, 1(5):349–360, 2014.

[33] J. M. Holton. A beginner's guide to radiation damage. *Journal of Synchrotron Radiation*, 16(2):133–142, 2009.

[34] P. Hopper, S. C. Harrison, and R. T. Sauer. Structure of tomato bushy stunt virus: V. Coat protein sequence determination and its structural implications. *Journal of Molecular Biology*, 177(4):701–713, 1984.

[35] M.R. Howells, T. Beetz, H.N. Chapman, C. Cui, J.M. Holton, C.J. Jacobsen, J. Kirz, E. Lima, S. Marchesini, H. Miao, D. Sayre, D.A. Shapiro, J.C.H. Spence, and D. Starodub. An assessment of the resolution limitation due to radiation-damage in X-ray diffraction microscopy. *Journal of Electron Spectroscopy and Related Phenomena*, 170(1):4–12, 2009.

[36] J. H. Hubbell. Review of photon interaction cross section data in the medical and biological context. *Physics in Medicine & Biology*, 44(1):R1–R22, 1999.

[37] W. Kabsch. Integration, scaling, space-group assignment and post-refinement. *Acta Crystallographica Section D*, 66(2):133–144, 2010.

[38] P. A. Karplus and K. Diederichs. Linking Crystallographic Model and Data Quality. *Science*, 336(6084):1030–1033, 2012.

[39] R. A. Kirian, X. Wang, U. Weierstall, K. E. Schmidt, J. C. H. Spence, M. Hunter, P. Fromme, T. A. White, H. N. Chapman, and J. Holton. Femtosecond protein nanocrystallography—data analysis methods. *Optics Express*, 18(6):5713–5723, 2010.

[40] G. Kovácsová, M. L. Grünbein, M. Kloos, T. R. M. Barends, R. Schlesinger, J. Heberle, W. Kabsch, R. L. Shoeman, R. B. Doak, and I. Schlichting. Viscous hydrophilic injection matrices for serial crystallography. *IUCrJ*, 4(4):400–410, 2017.

[41] T.-Y. Lan, J. L. Wierman, M. W. Tate, H. T. Philipp, V. Elser, and S. M. Gruner. Reconstructing three-dimensional protein crystal intensities from sparse unoriented two-axis X-ray diffraction patterns. *Journal of Applied Crystallography*, 50(4):985–993, 2017.

[42] T.-Y. Lan, J. L. Wierman, M. W. Tate, H. T. Philipp, J. M. Martin-Garcia, L. Zhu, D. Kissick, P. Fromme, R. F. Fischetti, W. Liu, V. Elser, and S. M. Gruner. Solving protein structure from sparse serial microcrystal diffraction data at a storage ring synchrotron source. *IUCrJ*, in press.

[43] X. Li, P. Mooney, S. Zheng, C. R. Booth, M. B. Braunfeld, S. Gubbens, D. A. Agard, and Y. Cheng. Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. *Nature Methods*, 10:584–590, 2013.

[44] M. Liang, G. J. Williams, M. Messerschmidt, M. M. Seibert, P. A. Montanez, M. Hayes, D. Milathianaki, A. Aquila, M. S. Hunter, J. E. Koglin, D. W. Schafer, S. Guillet, A. Busse, R. Bergan, W. Olson, K. Fox, N. Stewart, R. Curtis, A. A. Miahnahri, and S. Boutet. The Coherent X-ray Imaging instrument at the Linac Coherent Light Source. *Journal of Synchrotron Radiation*, 22(3):514–519, 2015.

[45] N. D. Loh, M. J. Bogan, V. Elser, A. Barty, S. Boutet, S. Bajt, J. Hajdu, T. Ekeberg, F. R. N. C. Maia, J. Schulz, M. M. Seibert, B. Iwan, N. Timneanu, S. Marchesini, I. Schlichting, R. L. Shoeman, L. Lomb, M. Frank, M. Liang, and H. N. Chapman. Cryptotomography: Reconstructing 3D Fourier Intensities from Randomly Oriented Single-Shot Diffraction Patterns. *Physical Review Letter*, 104:225501, 2010.

[46] N. D. Loh and V. Elser. Reconstruction algorithm for single-particle diffraction imaging experiments. *Physical Review E*, 80(2):026705, 2009.

[47] F. R. N. C. Maia. The Coherent X-ray Imaging Data Bank. *Nature Methods*, 9:854–855, 2012.

[48] J. M. Martin-Garcia, C. E. Conrad, G. Nelson, N. Stander, N. A. Zatsepin, J. Zook, L. Zhu, J. Geiger, E. Chun, D. Kissick, M. C. Hilgart, C. Ogata, A. Ishchenko, N. Nagaratnam, S. Roy-Chowdhury, J. Coe, G. Subramanian, A. Schaffer, D. James, G. Ketwala, N. Venugopalan, S. Xu, S. Corcoran, D. Ferguson, U. Weierstall, J. C. H. Spence, V. Cherezov, P. Fromme, R. F. Fischetti, and W. Liu. Serial millisecond crystallography of membrane and soluble protein microcrystals using synchrotron radiation. *IUCrJ*, 4:439–454, 2017.

[49] A. Munke, J. Andreasson, A. Aquila, S. Awel, K. Ayyer, A. Barty, R. J. Bean, P. Berntsen, J. Bielecki, S. Boutet, M. Bucher, H. N. Chapman, B. J. Daurer, H. DeMirci, V. Elser, P. Fromme, J. Hajdu, M. F. Hantke, A. Higashiura, B. G. Hogue, A. Hosseinizadeh, Y. Kim, R. A. Kirian, H. K. N. Reddy, T.-Y. Lan, D. S. D. Larsson, H. Liu, N. D. Loh, F. R. N. C. Maia, A. P. Mancuso, K. Mühlig, A. Nakagawa, D. Nam, G. Nelson, C. Nettelblad, K. Okamoto, A. Ourmazd, M. Rose, G. van der Schot, P. Schwander, M. M. Seibert, J. A. Sellberg, R. G. Sierra, C. Song, M. Svenda, N. Timneanu, I. A. Vartanyants, D. Westphal, M. O. Wiedorn, G. J. Williams, P. L. Xavier, C. H. Yoon, and J. Zook. Coherent diffraction of single Rice Dwarf virus particles using hard X-rays at the Linac Coherent Light Source. *Scientific Data*, 3(160064), 2016.

[50] G. N. Murshudov, P. Skubák, A. A. Lebedev, N. S. Pannu, R. A. Steiner, R. A. Nicholls, M. D. Winn, F. Long, and A. A. Vagin. *REFMAC*5 for the refinement of macromolecular crystal structures. *Acta Crystallographica Section D*, 67(4):355–367, 2011.

[51] A. Nakagawa, N. Miyazaki, J. Taka, H. Naitow, A. Ogawa, Z. Fujimoto, H. Mizuno, T. Higashi, Y. Watanabe, T. Omura, R. H. Cheng, and T. Tsukihara. The atomic structure of rice dwarf virus reveals the self-assembly mechanism of component proteins. *Structure*, 11(10):1227–1238, 2003.

[52] A. H. Narten and H. A. Levy. Liquid Water: Molecular Correlation Functions from X-Ray Diffraction. *The Journal of Chemical Physics*, 55(5):2263–2269, 1971.

[53] R. M. Neal and G. E. Hinton. A View of the EM Algorithm that Justifies Incremental, Sparse, and Other Variants. In *Learning in Graphical Models*, pages 355–368. Springer, 1999.

[54] R. Neutze, R. Wouts, D. van der Spoel, E. Weckert, and J. Hajdu. Potential for biomolecular imaging with femtosecond X-ray pulses. *Nature*, 406(6797):752–757, 2000.

[55] P. Nogly, D. James, D. Wang, T. A. White, N. Zatsepin, A. Shilova, G. Nelson, H. Liu, L. Johansson, M. Heymann, K. Jaeger, M. Metz, C. Wickstrand, W. Wu, P. Båth, P. Berntsen, D. Oberthuer, V. Panneels, V. Cherezov, H. N. Chapman, G. Schertler, R. Neutze, J. C. H. Spence, I. Moraes, M. Burghammer, J. Standfuss, and U. Weierstall. Lipidic cubic phase serial millisecond crystallography using synchrotron radiation. *IUCrJ*, 2(2):168–176, 2015.

[56] R. L. Owen, D. Axford, J. E. Nettleship, R. J. Owens, J. I. Robinson, A. W. Morgan, A. S. Doré, G. Lebon, C. G. Tate, E. E. Fry, J. Ren, D. I. Stuart, and G. Evans. Outrunning free radicals in room-temperature macromolecular crystallography. *Acta Crystallographica Section D*, 68(7):810–818, 2012.

[57] R. L. Owen, D. Axford, D. A. Sherrell, A. Kuo, O. P. Ernst, E. C. Schulz, R. J. D. Miller, and H. M. Mueller-Werkmeister. Low-dose fixed-target serial synchrotron crystallography. *Acta Crystallographica Section D*, 73(4):373–378, 2017.

[58] H. T. Philipp, K. Ayyer, M. W. Tate, V. Elser, and S. M. Gruner. Solving structure with sparse, randomly-oriented x-ray data. *Opt. Express*, 20(12):13129–13137, 2012.

[59] H. T. Philipp, L. J. Koerner, M. S. Hromalik, M. W. Tate, and S. M. Gruner. Femtosecond radiation experiment detector for X-ray Free-Electron Laser (XFEL)

coherent X-ray imaging. In *Nuclear Science Symposium Conference Record, 2008. NSS '08. IEEE*, pages 1567–1571, 2008.

[60] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes: The Art of Scientific Computing*, chapter 10. Cambridge University Press, 3rd edition, 2007.

[61] H. K. N. Reddy, C. H. Yoon, A. Aquila, S. Awel, K. Ayyer, A. Barty, P. Berntsen, J. Bielecki, S. Bobkov, M. Bucher, G. A. Carini, S. Carron, H. N. Chapman, B. J. Daurer, H. DeMirci, T. Ekeberg, P. Fromme, J. Hajdu, M. F. Hanke, P. Hart, B. G. Hogue, A. Hosseinizadeh, Y. Kim, R. A. Kirian, R. P. Kurta, D. S. D. Larsson, N. D. Loh, F. R. N. C. Maia, A. P. Mancuso, K. Mühlig, A. Munke, D. Nam, C. Nettelblad, A. Ourmazd, M. Rose, P. Schwander, M. M. Seibert, J. A. Sellberg, C. Song, J. C. H. Spence, M. Svenda, G. van der Schot, I. A. Vartanyants, G. J. Williams, and P. L. Xavier. Coherent soft X-ray diffraction imaging of coliphage PR772 at the Linac coherent light source. *Scientific Data*, 4(170079), 2017.

[62] P. Roedig, R. Duman, J. Sanchez-Weatherby, I. Vartiainen, A. Burkhardt, M. Warmer, C. David, A. Wagner, and A. Meents. Room-temperature macromolecular crystallography using a micro-patterned silicon chip with minimal background scattering. *Journal of Applied Crystallography*, 49(3):968–975, 2016.

[63] M. G. Rossmann and E. Arnold, editors. *International tables for crystallography*, volume F. Chester: International Union of Crystallography, 1st online edition, 2006.

[64] D. Sayre. Some implications of a theorem due to Shannon. *Acta Crystallographica*, 5(6):843, 1952.

[65] C. E. Shannon. Communication in the Presence of Noise. *Proceedings of the IRE*, 37(1):10–21, 1949.

[66] S. C. Shoemaker and N. Ando. X-rays in the Cryo-Electron Microscopy Era: Structural Biology's Dynamic Future. *Biochemistry*, 57(3):277–285, 2018.

[67] L. B. Skinner, C. Huang, D. Schlesinger, L. G. M. Pettersson, A. Nilsson, and C. J. Benmore. Benchmark oxygen-oxygen pair-distribution function of ambient water from x-ray diffraction measurements with a wide Q-range. *The Journal of Chemical Physics*, 138(7):074506, 2013.

[68] J. C. H. Spence. XFELs for structure and dynamics in biology. *IUCrJ*, 4(4):322–339, 2017.

[69] F. Stellato, D. Oberthür, M. Liang, R. Bean, C. Gati, O. Yefanov, A. Barty, A. Burkhardt, P. Fischer, L. Galli, R. A. Kirian, J. Meyer, S. Panneerselvam, C. H. Yoon, F. Chervinskii, E. Speller, T. A. White, C. Betzel, A. Meents, and H. N. Chapman. Room-temperature macromolecular serial crystallography using synchrotron radiation. *IUCrJ*, 1(4):204–212, 2014.

[70] I. Steller, R. Bolotovsky, and M. G. Rossmann. An Algorithm for Automatic Indexing of Oscillation Images using Fourier Analysis. *Journal of Applied Crystallography*, 30(6):1036–1040, 1997.

[71] M. W. Tate, D. Chamberlain, K. S. Green, H. T. Philipp, P. Purohit, C. Strohman, and S. M. Gruner. A Medium-Format, Mixed-Mode Pixel Array Detector for Kilohertz X-ray Imaging. *Journal of Physics: Conference Series*, 425(6):062004, 2013.

[72] P. Thibault. *Algorithmic methods in diffraction microscopy*. PhD thesis, Cornell University, 2007.

[73] A. Vagin and A. Teplyakov. Molecular replacement with *MOLREP*. *Acta Crystallographica Section D*, 66(1), 2010.

[74] K. Valegard, J. B. Murray, N. J. Stonehouse, S. van den Worm, P. G. Stockley, and L. Liljas. The three-dimensional structures of two complexes between recombinant MS2 capsids and RNA operator fragments reveal sequence-specific protein-RNA interactions. *Journal of Molecular Biology*, 270(5):724–738, 1997.

[75] M. C. Vaney, S. Maignan, M. Riès-Kautt, and A. Ducruix. High-Resolution Structure (1.33 Å) of a HEW Lysozyme Tetragonal Crystal Grown in the APCF Apparatus. Data and Structural Comparison with a Crystal Grown under Microgravity from SpaceHab-01 Mission. *Acta Crystallographica Section D*, 52(3):505–517, 1996.

[76] U. Weierstall, D. James, C. Wang, T. A. White, D. Wang, W. Liu, J. C. H. Spence, R. B. Doak, G. Nelson, P. Fromme, R. Fromme, I. Grotjohann, C. Kupitz, N. A. Zatsepin, H. Liu, S. Basu, D. Wacker, G. W. Han, V. Katritch, S. Boutet, M. Messerschmidt, G. J. Williams, J. E. Koglin, M. M. Seibert, M. Klinker, C. Gati, R. L. Shoeman, A. Barty, H. N. Chapman, R. A. Kirian, K. R. Beyerlein, R. C. Stevens, D. Li, S. T. A. Shah, N. Howe, M. Caffrey, and V. Cherezov. Lipidic cubic phase injector facilitates membrane protein serial femtosecond crystallography. *Nature Communication*, 5(3309), 2014.

[77] T. A. White, R. A. Kirian, A. V. Martin, A. Aquila, K. Nass, A. Barty, and H. N.

Chapman. *CrystFEL*: a software suite for snapshot serial crystallography. *Journal of Applied Crystallography*, 45(2):335–341, 2012.

[78] M. O. Wiedorn, S. Awel, A. J. Morgan, M. Barthelmess, R. Bean, K. R. Beyerlein, L. M. G. Chavas, N. Eckerskorn, H. Fleckenstein, M. Heymann, D. A. Horke, J. Knoška, V. Mariani, D. Oberthür, N. Roth, O. Yefanov, A. Barty, S. Bajt, J. Küpper, A. V. Rode, R. A. Kirian, and H. N. Chapman. Post-sample aperture for low background diffraction experiments at X-ray free-electron lasers. *Journal of Synchrotron Radiation*, 24(6):1296–1298, 2017.

[79] J. L. Wierman, T.-Y. Lan, M. W. Tate, H. T. Philipp, V. Elser, and S. M. Gruner. Protein crystal structure from non-oriented, single-axis sparse X-ray data. *IUCrJ*, 3(1):43–50, 2016.

[80] M. D. Winn, C. C. Ballard, K. D. Cowtan, E. J. Dodson, P. Emsley, P. R. Evans, R. M. Keegan, E. B. Krissinel, A. G. W. Leslie, A. McCoy, S. J. McNicholas, G. N. Murshudov, N. S. Pannu, E. A. Potterton, H. R. Powell, R. J. Read, A. Vagin, and K. S. Wilson. Overview of the *CCP*4 suite and current developments. *Acta Crystallographica Section D*, 67(4):235–242, 2011.

[81] Z. Zhang, N. K. Sauter, H. van den Bedem, G. Snell, and A. M. Deacon. Automated diffraction image analysis and spot searching for high-throughput crystal screening. *Journal of Applied Crystallography*, 39(1):112–119, 2006.